EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume XXVI

1966

BOX 6907, COLLEGE STATION, DURHAM, N. C. 27708

EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

VOLUME TWENTY-SIX, NUMBER FOUR, WINTER, 1966

INDEX FOR VOLUME 26

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of October 23, 1962; Section 4369, Title 39, United States Code)

1. DATE OF FILING
   September 14, 1966
2. TITLE OF PUBLICATION
   Educational and Psychological Measurement
3. FREQUENCY OF ISSUE
   Quarterly
4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, zip code)*
   2901 Byrdhill Road, Richmond, Virginia 23228
5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
   3121 Cheek Road, Durham, N. C. 27703
6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR
   PUBLISHER *(Name and address)*
   G. Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708
   EDITOR *(Name and address)*
   G. Frederic Kuder, Box 6907 College Station, Durham, N. C. 27708
   MANAGING EDITOR *(Name and address)*
   Geraldine R. Thomas, 3121 Cheek Road, Durham, N. C. 27703
7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
G. Frederic Kuder (Owner)	**Box 6907 College Station, Durham, N. C. 27708**

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*
   None
9. THIS ITEM MUST BE COMPLETED FOR ALL PUBLICATIONS EXCEPT THOSE WHICH DO NOT CARRY ADVERTISING OTHER THAN THE PUBLISHER'S OWN AND WHICH ARE NAMED IN SECTIONS 132.231, 132.232, AND 132.233, POSTAL MANUAL *(Sections 4355a, 4355b, and 4356 of Title 39, United States Code)*

I certify that the statements made by me above are correct and complete.

(Signature of editor, publisher, business manager, or owner)
Geraldine R. Thomas, Managing Editor

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

MULTIPLE RANK ORDER TYPAL ANALYSIS FOR THE ISOLATION OF INDEPENDENT TYPES

LOUIS L. McQUITTY
Michigan State University

Whether or not types are found to exist depends in part on how they are defined. Quantitative methods for the isolation of types imply definitions of types. In the search for types, it is therefore helpful to have several quantitative methods, each related to a reasonable definition of types. The several methods could then be applied to determine which yields the most fruitful types in terms of some specified criteria.

It would be still better if a single method were available which would both isolate and define all reasonable types reflected in a set of data. Pending availability of the latter alternative it is helpful to proceed with development of methods under the former alternative, and this course of action might itself lead to a realization of the latter alternative.

We have already described and illustrated Multiple Rank Order Typal Analysis for the Isolation of *Intersecting Types* (McQuitty, 1965). The present paper develops and illustrates an associated version of Rank Order Typal Analysis for the Isolation of *Independent Types.*

Theory in the Development of Methods

Intersecting Types and Typologies

In our development of the version for the isolation of *intersecting types,* we applied several assumptions. We assumed that several, or even many, sets of pressures act on the developing person (institution or other object). Each set, or at least a few of

them, tends toward the development of types. Consequently, the individual is not a member of just one type but rather of many imperfect (incomplete) types. However, these types cannot be seen with the "unaided eye"; the several types have become intermingled to such an extent that they are hidden.

The version for isolating *intersecting types* serves to disentangle and display the several types. It is not, however, applied only to the single individual; it is applied to matrices of individuals and displays multiple typologies, showing each person's location in each typology and at least some of the significant characteristics which define him there.

Independent Types and Typologies

There is reason to believe that diverse pressures sometimes have more independent effects than are indicated above, at least in the case of some people. It is possible that each of several major pressures may have an effect on the developing personality and that each effect may be relatively independent of every other effect, thus producing relatively independent types and typologies. This kind of a result is implied in the personality descriptions of the mentally ill and to a lesser extent in the personalities of the normal by those whose theories developed out of psychopathology, primarily.

Simplicity of Structure

Another argument in favor of the *Independent* Version is the fact that it yields a simpler model; each characteristic of a person is used only once in classifying him as compared with the possibility of multiple usage in the case of the *Intersecting* Versions.

A Review of Rank Order Typal Analysis

"Rank Order Typal Analysis defines a type as a category of people of such a nature that everyone in the category is more like everyone else in the category than he is like anyone in any other category. . . .

"It is a simple matter to examine a matrix of interassociations between people in such a fashion as to isolate all of the categories which qualify as types. The task is accomplished by examining the matrix serially one column at a time.

"The first step is to take a matrix such as the one shown in Table 1 and convert it into a matrix of ranks as reported in Table 2

TABLE 1

Hypothetical Associations between People

	w	x	y	z
w	71	60	65	35
x	60	72	50	40
y	65	50	73	45
z	35	40	45	74

TABLE 2

Rank Order within the Columns from Table 1

	w	x	y	z
w	1	2	2	4
x	3	1	3	3
y	2	3	1	2
z	4	4	4	1

to show, for example, that the person most like x is x himself and then w, y, and z in that order.

"If x forms a type with any one person of Table 2, it must be with the Person w, who is second most like him (x being most like himself). In order to examine this possibility, we form a submatrix of x with w, Table 3, using the ranks from Table 2. Person x does not form a type with Person w; the submatrix contains a rank larger than the number of persons in the matrix. Since Person x does not form a type of two persons with the one person

TABLE 3

A Submatrix from Table 2

	w	x
w	1	2
x	3	1

TABLE 4

A Submatrix from Table 2

	w	x	y
w	1	2	2
x	3	1	3
y	2	3	1

second most like x (x being most like himself), Person x does not form a type of two persons with any other one person of Table 2.

"If x forms a type with any two persons of Table 2, it must be with the persons second and third most like x. These are w and y. A submatrix of w, x, and y is formed, Table 4. It contains no rank larger than 3. It proves, therefore, that x forms a type of three people with w and y and furthermore forms no other type of three people from those of Table 2.

"The analysis proceeds in this fashion until Column x is completed. In case of a 4 x 4 matrix, as in Table 2, a column has been completed when three variables have been included in a submatrix, as they have in Table 4; the analysis for x was completed in the above operations.

"In the general case, Column i of an nXn matrix has been completed when a submatrix of n—1 variables has derived from a study of Column i. If all of the variables of a matrix were included in a submatrix, they would by necessity meet the above criterion but the criterion would no longer operate as a test of the presence of a type.

"Other columns of the matrix are analyzed in a similar fashion. In order to save time, steps should be taken to avoid duplications in the types isolated. For example, if Person x forms a type of three persons with Persons w and y, then each of these latter persons will yield the same type; they, therefore, need not be examined for this possibility.

"In case of ties of more than two indices, it is helpful to follow a method of assigning ranks which deviate from the usual method. In the new method, if Rank 7, for example, is followed by three tied indices such as 23, 23, 23, and then a larger index, say of 25, the three tied indices are assigned a rank of 10—the highest of the ranks in the series required for the three tied entries. The next entry is assigned a rank of 11.

"In the general case, tied indices are each assigned the highest rank of the series required to cover the tied values. This approach prevents, for example, three tied cases from first forming three types of two cases each and then a type of three cases when all three cases are equally alike.

"Once the data have been analyzed in the fashion outlined above, all possible types have been isolated, i.e., all possible types which

derive from all of the data, with every item equal in weight to every other item; other types can possibly be isolated in terms of reduced sets of data as will now be shown" (McQuitty, 1965).

The Isolation of Multiple, Independent Types

In order to illustrate Multiple Rank Order Typal Analysis for the Isolation of Independent Types we applied it to the classification of eight industrial companies in terms of union-management characteristics. These data were again chosen for two reasons: (a) they are a small set of highly structured data which can be economically analyzed and displayed, and (b) the results can be compared concisely with those of earlier analyses of the same data by other methods.

The matrix of interassociation between the eight companies is shown in Table 5 below.

TABLE 5

*Agreement Scores between Companies**

	A	B	C	D	E	F	G	H
A	32	29	16	16	14	6	11	7
B	29	32	17	17	13	6	8	10
C	16	17	32	26	10	8	9	13
D	16	17	26	32	10	12	11	11
E	14	13	10	10	32	21	17	13
F	6	6	8	12	21	32	19	17
G	11	8	9	11	17	19	32	24
H	7	10	13	11	13	17	24	32

* Data from McQuitty, 1954

The first step of both the *Intersecting* and the *Independent* Versions is to apply a Rank Order Typal Analysis (McQuitty, 1963) as outlined above.

The Rank Order Typal Analysis yielded the following results: Companies A and B, two construction companies, first joined to form a type, followed in order by C and D, two trucking companies; G and H, two garment manufacturing companies; and then by E and F, a grain processing company and a products manufacturing company. In the final operation, two types, AB and CD, joined to form a new type, ABCD; no other combinations of companies qualified as a type under our above definition.

The steps unique to the *Independent* Version were applied as

follows: the most inclusive types were defined. They are the minimum number of types required to include all of the companies which classify into types at this stage; in this case they are Types ABCD, EF, and GH; Types AB and CD were omitted because they are included in Type ABCD. For each inclusive type, the characteristics common to the members of that type were first specified and then they were subtracted out of the original description of each of the individual companies. For example, Companies A, B, C, and D agreed on 13 of the 32 original, dichotomous items, and the subtraction left each of these four companies described in terms of but 19 items. As a result, each of these companies could no longer agree with any company in terms of the items common to Type ABCD.

Analogously, the common characteristics were selected out of the description of the two companies for each of the other two inclusive types, Types EF and GH.

A new matrix of agreement scores was prepared between all of the 8 companies in terms of the items which remained for each individual company.

A Rank Order Typal Analysis was performed on this matrix. It yielded Types AB and CD; no other combinations qualified as types. These Types, AB and CD, had already appeared on the way to yielding Type ABCD.

The method proceeds with Types AB and CD, just as it would with any other "inclusive" types (as defined above).

The common characteristics of each Type AB and Type CD were subtracted from each A and B, on the one hand, and C and D, on the other.

A new matrix of agreement scores was formed, just the same as the last one except for the omissions of items in Companies A, B, C, and D as specified in the above paragraph.

The last matrix was analyzed by Rank Order Typal Analysis and failed to yield any types. The analysis was thus completed.

Versions in the Search for Independent Types

There are versions in the method of searching for *independent types*. We could have used *first-order types* rather than *inclusive types* in the operation where we eliminate characteristics common to the members of the types isolated.

In this study, the *first-order types* of the first application of Rank Order Typal Analysis are Types AB, CD, EF, and GH; they are the lowest order of types within the typology which includes every company with one or more other companies and does not include any company more than once.

If we had used the *first-order types,* AB, CD, EF, and GH, rather than the *inclusive types,* ABCD, EF, and GH, the first elimination would have produced the final matrix.

The above point is clarified by a brief review of steps in the present analysis. We eliminated what was common to the members of Type ABCD. The subsequent analysis produced again Types AB and CD, this time as *inclusive types,* and we then eliminated what was still common on the one hand to the members of Type AB and on the other hand what was still common to the members of Type CD.

The use of *first-order types* would have eliminated one step in the present study. This result does not ensue because of the nature of the method alone, but because of the nature of the method in relation to the data.

The use of *inclusive types* as contrasted with *first-order types* provides for the isolation of a typology (or typologies) which might otherwise remain hidden. However, only the *inclusive types* of the first analysis would of necessity be independent of the types of the subsequent analyses; the *first-order types* of the first analysis would likely intersect with types of the subsequent analyses; they would not, however, of necessity intersect with them.

The use of *first-order types* would require that all types of the first analysis be independent of all types of subsequent analyses.

The two analyses yielded the results shown in Figures 1 and 2.

Figure 1 reports the exact number of characteristics that the members of each type have in common. Furthermore, it reveals the

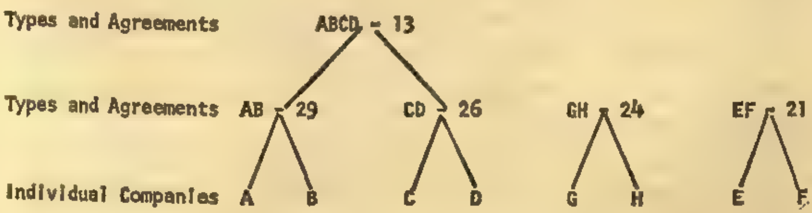


Figure 1. First Analysis

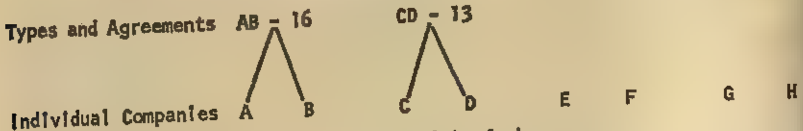


Figure 2. Second Analysis

number that the members of each AB and CD have in common over and above those which all four companies have in common as members of Type ABCD; they are 16 and 13, respectively (29–13=16 for Type AB and 26–13=13 for Type CD). This latter fact is reported also in Figure 2, but Figure 2 still adds information not revealed by Figure 1; it shows that the data of this study reveal only one system of *independent types*, in terms of the stringent definition of types used in the above analysis.

Interpretation

It is helpful to contrast the above analysis in search of *independent types* with those of the same data in search of *intersecting types* (McQuitty, 1965).

In addition to the typology shown in Figure 1, the search for *intersecting types* yielded the typology of Figure 3.

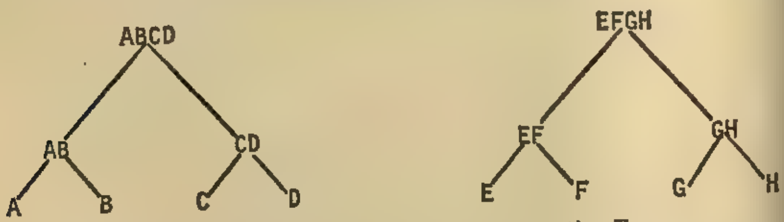


Figure 3. A Typology for a Search for *Intersecting Types*

The results of Figure 3 were obtained if we performed a Rank Order Typal Analysis using the characteristics of either Type CD, EF, or GH, exclusively, but not if we used either all of the characteristics of the original study or those of Type AB exclusively. In these latter two cases, we obtained the typology of Figure 1.

The agreement scores of the types in Figure 3 are not reported because they vary depending on whether the set of characteristics used was from Type CD, EF, or GH.

The present set of data reflects two similar configurations of *intersecting* typologies but only one system of *independent types.* The substantive implications of this very limited study are that

various pressures do in fact act on institutions to yield more than one typology and that the pressures interact to yield *intersecting* rather than *independent types.*

The substantive implications of this study need to be investigated further with much more comprehensive sets of data in relation to people, institutions, and other objects in many cultures. The contributions of this and related papers are in the methodologies which they provide for the substantive studies.

Summary

This paper assumes that various sets of pressures act on individuals, (institutions and other objects) as they develop. As a consequence, each of the several individuals tends to develop more than one type, and a sample of individuals reflects more than one typology. The multiple existence of types obscures them from unaided observation.

Quantitative methods are needed to help separate out and display the various types and typologies. In developing such methods, it is possible to assume either that the types (and the typologies) are independent, or that they are intersecting.

In this paper, it is assumed that the multiple types (and multiple typologies) are *independent*, and a quantitative method to assist in isolating them is developed and illustrated with empirical data.

REFERENCES

McQuitty, L. L. "Pattern-Analysis: A Statistical Method for the Study of Types." In W. E. Chalmers, M. K. Chandler, L. L. McQuitty, R. Stagner, D. E. Wray, and M. Derber (Editors). *Labor Management Relations in Illini City*, Volume II. Champaign, Illinois: Institute of Labor and Industrial Relations, University of Illinois, 1954.

McQuitty, L. L. "Rank Order Typal Analysis." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 55–61.

McQuitty, L. L. "A Conjunction of Rank Order Typal Analysis and Item Selection." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXV (1965), 949–961.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

KUDER-RICHARDSON RELIABILITIES OF CLASSROOM TESTS

EDWARD E. CURETON
University of Tennessee

Classroom tests often contain substantial numbers of items so easy that everyone or almost everyone gets them right. The purpose of the present note is to show the effects of this situation on estimates of the reliabilities of such tests by the Kuder-Richardson formulas 20 and 21. Ordinarily the too-easy items are merely considered "dead wood": they take up the time of the examinees but are not supposed to contribute either positively or negatively to the reliability of the test.

We will use an extreme case as an example. Suppose we have initially a test of k "good" items; "good" in the sense that they all measure the same trait or ability, apart from error, and that every one of them has positive and significant discrimination. To this test we add k additional items, all so easy that every examinee gives the right answer to every one of them. We assume that every item is scored 1 for right and 0 for wrong. For the initial set, we have

$$KR_{20} = \frac{k}{k-1}\left[1 - \frac{Spq}{\sigma_t^2}\right],$$

where S is a summation over the k items and σ_t^2 is the variance of the total scores. The added items will not change the variance of the total scores, and pq will equal 0 for every one of them, so the expression in brackets will remain unchanged. For the augmented test we then have,

$$KR_{20} = \frac{2k}{2k-1}\left[1 - \frac{Spq}{\sigma_t^2}\right].$$

The estimate of reliability will be reduced slightly but not seriously.
Thus, if $k = 25$, $k/(k-1) = 1.04$, and $2k/(2k-1) = 1.02$.
Also for the initial set of k items,

$$KR_{21} = \frac{k}{k-1}\left[1 - \frac{\bar{X} - \bar{X}^2/k}{\sigma_t^2}\right]$$

Again the added items will not change σ_t^2, but $\overline{X}$ will be increased by k. For the augmented test we than have,

$$KR_{21} = \frac{2k}{2k-1}\left[1 - \frac{\bar{X} + k - (\bar{X} + k)^2/2k}{\sigma_t^2}\right]$$

$$= \frac{2k}{2k-1}\left[1 - \frac{\bar{X} - \bar{X}^2/k + (k - \bar{X})^2/2k}{\sigma_t^2}\right]$$

The added term is intrinsically positive, so the reliability of the augmented test will be still further underestimated by KR_{21}. Thus if $k = 25$, $\overline{X} = 15$, and $\sigma_t^2 = 4$, KR_{21} will give .65 for the initial test and .51 for the augmented test.

It follows that KR_{21} should not be recommended as a short formula for estimating the reliabilities of classroom tests, or in general, of any tests whose nonfunctional items have not been eliminated in advance by item analysis.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 1, 1966

A COMPARISON OF 57 MMPI SCALES AND 57 EXPERIMENTAL SCALES MATCHED WITH THE MMPI SCALES IN TERMS OF ITEM SOCIAL DESIRABILITY SCALE VALUES AND PROBABILITIES OF ENDORSEMENT[1]

ALLEN L. EDWARDS[2]
University of Washington

IT has been shown (Edwards and Diers, 1962; Edwards, Diers, and Walker, 1962; Edwards and Heathers, 1962) that first factor loadings of Minnesota Multiphasic Personality Inventory (MMPI) scales are highly correlated with the zero order correlations of the scales with Edwards' (1957) Social Desirability (SD) scale. The SD scale was designed to measure a general personality trait, namely, the tendency to give socially desirable responses in self-description under the standard instructions ordinarily used in administering personality scales and inventories. A socially desirable response is defined as a True response to an item with a socially desirable scale value or as a False response to an item with a socially undesirable scale value. All of the items in the SD scale are keyed for socially desirable responses.

If the trait keyed responses to the items in a trait scale are also socially desirable responses, then the trait keying and the social desirability keying are completely confounded. A high score on the trait scale may be obtained by a subject who responds to the items in terms of the trait which the scale was designed to measure,

[1] This research was supported in part by Research Grant MH 04075 from the National Institute of Mental Health, United States Public Health Service.

[2] I am indebted to Dr. Benard H. Taylor for assistance in the collection of the test records used in this study; to Dr. James A. Walsh for selecting the items in the Experimental Personality Inventory; and to Dr. Walsh and Alan J. Klockars for supervising the runs on the computer.

but a high score may also be obtained by a subject who has a strong tendency to give socially desirable responses. If the tendency to give socially desirable responses operates with respect to personality items in general and not just with respect to the items in the SD scale, then scores on trait scales which have a large proportion of items keyed for socially desirable responses should be positively correlated with scores on the SD scale. Similar considerations apply to trait scales which have a small proportion of items keyed for socially desirable responses and, consequently, a large proportion of items keyed for socially undesirable responses. If the tendency to give socially desirable responses is operating, then, in this instance, high scores on the SD scale should be associated with low scores on the trait scale and the correlation between the trait scale and the SD scale should be negative. The fact that both the correlations of MMPI scales with the SD scale and the first factor loadings of the scales have been found to be highly correlated with the proportion of items keyed for socially desirable responses in the scales (Edwards, 1961; Edwards and Diers, 1962; Edwards, Diers, and Walker, 1962; Edwards and Heathers, 1962) offers fairly convincing evidence that responses to items in MMPI scales are influenced by social desirability tendencies.

The present study was undertaken to determine whether the relationship observed between the proportion of items keyed for socially desirable responses in MMPI scales and the correlations of these scales with the SD scale is restricted to scales composed of MMPI items. More specifically, if we build experimental scales consisting of non-MMPI items but such that each experimental scale has certain structural properties in common with a corresponding MMPI scale, will the correlations of the experimental scales with the SD scale be proportional to the correlations of the corresponding MMPI scales with the SD scale? Furthermore, will the internal consistency of the experimental scale be proportional to the internal consistency of the corresponding MMPI scale? And will the correlation between the experimental scale and the corresponding MMPI scale be related to the proportion of items keyed for socially desirable responses in the scales?

MMPI scales have been constructed by a variety of techniques. Some scales were developed by finding those items in the MMPI

which differentiated between a criterion and a control group. Others were constructed using some form of internal consistency analysis. Still others were developed on a rational or construct basis. Presumably, each MMPI scale was developed to measure some trait of interest and, presumably, each scale has some degree of predictive, content, construct, or concurrent validity. The experimental scales constructed for the present study, on the other hand, cannot be said to possess any validity of any kind in any degree. The basis for this statement is that the experimental scales have never been scored before and the items in each of the scales were not selected on the basis of content or by means of any kind of internal consistency analysis.

The experimental scales, however, do have some structural properties in common with MMPI scales. First of all, for each MMPI scale there is a corresponding experimental scale scored by the same scoring stencil as the MMPI scale. The corresponding MMPI and experimental scales have the following properties in common: (1) each scale contains the same number of items; (2) the location of the items in the experimental scale in a set of 566 items is the same as the location of the items in the MMPI scale in a set of 566 items; (3) the corresponding items in the experimental scale and the MMPI scale have approximately the same social desirability scale values and also approximately the same probabilities of being answered True; and (4) the keying of the items in the experimental scale is exactly the same as the *trait* keying of the items in the MMPI scale. Thus, it follows that both scales have the same proportion of items keyed for socially desirable responses and that the intensity of the social desirability keying of the items in the corresponding experimental and MMPI scale is approximately the same.

Method and Procedure

Messick and Jackson (1961) have provided social desirability scale values (SDSVs), based upon the judgments of college students, for each of the 566 items in the MMPI. Goldberg and Rorer (1963) give the probabilities of a True response, P(T), based upon the responses of college students, for each of the MMPI items. Edwards (1963) has obtained SDSVs and P(T)s for each of the items in his pool of 2,824 experimental personality statements. The

2,824 experimental items were sorted according to their SDSVs. Then, using the SDSV of each item in the MMPI, a corresponding experimental item was selected from the pool of 2,824 items with approximately the same SDSV and P(T) as the MMPI item and without regard to the content of the experimental item. For example, the first three items in the MMPI are: (1) I like mechanics magazines; (2) I have a good appetite; and (3) I wake up fresh and rested most mornings. The corresponding items in the Experimental Personality Inventory (EPI) are: (1) I value logic above personal feelings; (2) I am not resistant to new ideas; and (3) I will keep at a problem long after others have given it up as hopeless. The SDSVs and P(T)s of these paired items are:

	MMPI		EPI	
Item	SDSV	P(T)	SDSV	P(T)
1.	5.48	.65	5.60	.58
2.	6.78	.96	6.78	.96
3.	7.75	.46	7.65	.46

In the same manner, each item in the EPI was matched with a corresponding item in the MMPI in terms of SDSV and P(T). No consideration was given to the content of the items selected for the EPI.

The MMPI was administered to a group of male college students. Approximately a week later the EPI was administered. Test records for both the MMPI and the EPI were obtained from 138 male students. Using the scoring stencils for each of 57 MMPI scales, scores on the 57 MMPI scales were obtained for each student, based upon his responses to the MMPI items.[3] The same scoring stencils were then applied to the student's EPI answer sheet to obtain corresponding scores on the 57 EPI scales, based upon his responses to the EPI items. Scores on the 114 scales were intercorrelated and factor analyzed by the method of principal components. Two factors were extracted. The first factor accounted for 37 percent of the total variance and the second factor accounted for 10 percent of the total variance. These percentages are comparable to those obtained when only the MMPI scales are fac-

[3] The code names of the MMPI scales are as follows: *A, Ac, Ad, Ai, At, B, Ca, Cn, D, Dn, Do, D-O, D-S, Dy, Eo, Es, F, Fm, Ho, Hs, Hy, Hy-O, Hy-S, Ie, Im, K, L, Lp, Ma, Ma-O, Ma-S, Mf-m, Mp, Ne, No, Nu, Or, Pa, Pa-O, Pa-S, Pd, Pd-O, Pd-S, Pn, Pr, Pt, Pv, R, Re, Rp, Sc, Si, SD, Sp, St, Tt, Sd.*

tor analyzed (Edwards, Diers, and Walker, 1962). Of primary interest are the first factor loadings of the scales, since it is the first factor of the MMPI which has been interpreted as a social desirability factor.

The original SD scale is composed of 39 MMPI items and this scale will be referred to as SD_1. The corresponding EPI SD scale will be referred to as SD_2. The correlations of the MMPI scales with scores on both SD_1 and SD_2 were obtained. The correlations between the MMPI scales and SD_2 cannot be attributed to item overlap because SD_2 contains no items in common with any of the MMPI scales. The corresponding correlations between the EPI scales with SD_1 and SD_2 were obtained. In this instance, the correlations between the EPI scales and SD_1 cannot be attributed to item overlap because SD_1 contains no items in common with any of the EPI scales.

The first factor loadings of each of the EPI and MMPI scales and the correlations of each of the scales with SD_1 and SD_2 represent dependent variables of interest. A common property of each of the corresponding EPI and MMPI scales is the proportion of items keyed for socially desirable responses and this is an independent variable of interest. We thus obtained a matrix with 57 rows, corresponding to the 57 scales, and 7 columns, corresponding to the first factor loadings of MMPI scales, the first factor loadings of EPI scales, the correlations of the MMPI scales with SD_1 and with SD_2, the correlations of the EPI scales with SD_1 and SD_2, and the proportion of items in the scales keyed for socially desirable responses. Correlations were then obtained between each of the column variables.

All of the variables described above, with the exception of the proportion of items keyed for socially desirable responses in the scales, are *signed* variables, that is, the first factor loadings of both EPI and MMPI scales are in some cases positive and in other cases negative and this is true also of the correlations of the scales with SD_1 and SD_2. All of these dependent signed variables, according to social desirability considerations, should be linearly related to one another and to the proportion of items keyed for socially desirable responses. There are other variables of interest however, which, according to social desirability considerations, should be non-linearly related to the signed first factor loadings and signed corre-

lations of the scales with the two SD scales but which should be linearly related to the *absolute* values of these variables.

For example, according to social desirability considerations, if we obtain Kuder-Richardson Formula 21 (K-R 21) values for each of the scales and plot these values against: (1) the signed first factor loadings, (2) the signed correlations of the scales with the two SD scales, or (3) the proportion of items keyed for socially desirable responses, we should expect to find higher K-R 21 values for scales which have values at both extremes of these three variables than for scales which have values in between the two extremes. Scales with middle values are scales which have: (1) low loadings on the first factor; (2) low correlations with the SD scales; and (3) a balance in their social desirability keying. Scales with any of these properties should have relatively lower K-R 21 values than scales which do not.

Similar considerations apply to the correlations between corresponding EPI and MMPI scales. All of these correlations are predicted to be positive in sign. But high values of the correlations should be associated with scales which have either a large or a small proportion of items keyed for socially desirable responses and low values should be associated with scales which have a better balance in their social desirability keying. Thus, it follows that high correlations between the EPI and MMPI scales should also be associated with scales which have either high positive or high negative first factor loadings and which have either high positive or high negative correlations with the SD scale.

For the reasons cited above, intercorrelations were also obtained between the K-R 21 values of the MMPI and EPI scales, the absolute values of the first factor loadings of the EPI and MMPI scales, the absolute values of the correlations of the scales with the two SD scales, the product of the first factor loadings of corresponding EPI and MMPI scales, and the correlations between corresponding EPI and MMPI scales.

Results and Discussion

Table 1 gives the intercorrelations between the signed first factor loadings of the EPI and MMPI scales, the signed correlations of the EPI and MMPI scales with the two SD scales, and the proportion of items keyed for socially desirable responses in the scales. It

is obvious that the first factor loadings of the MMPI scales could be predicted quite accurately ($r = .98$) from the factor loadings of the EPI scales.[4] It is also clear that the correlations of the EPI scales with SD_1 are proportional to the corresponding correlations of MMPI scales with SD_1, despite the fact that SD_1 contains no items in common with any of the EPI scales. Similarly, the correlations of the MMPI scales with SD_2 are proportional to the corresponding correlations of EPI scales with SD_2, despite the fact that SD_2 contains no items in common with any of the MMPI scales. Of interest also is the fact that the SD_1 correlations with the EPI scales are proportional to the SD_2 correlations with the MMPI scales.

TABLE 1[a]

Intercorrelations between the Reflected First Factor Loadings of 57 MMPI Scales (I), the Reflected First Factor Loadings of 57 EPI Scales (I′), the Correlations of the MMPI and EPI Scales with Two Social Desirability Scales (SD_1 and SD_2), and the Proportion of Items Keyed for Socially Desirable Responses in the Scales, P(SD). Means and Standard Deviations Are Given in the Last Two Rows

	MMPI I	EPI I′	MMPI SD_1	EPI SD_1	MMPI SD_2	EPI SD_2	P(SD)
I	—	.98	.99	.97	.99	.93	.90
I′		—	.98	.99	.97	.94	.90
MMPI: SD_1			—	.97	.99	.92	.91
EPI: SD_1				—	.96	.94	.90
MMPI: SD_2					—	.92	.93
EPI: SD_2						—	.87
$\bar{X}$	−.12	−.13	−.13	−.11	−.11	−.12	.39
s	.63	.58	.58	.44	.50	.54	.33

[a] The SD scale and other scales with a large proportion of items keyed for socially desirable responses have negative loadings on the unreflected first factor. The factor loadings were reflected so that these scales would have positive loadings on the reflected factor and scales with a small proportion of items keyed for socially desirable responses would have negative loadings.

The last column of Table 1 gives the correlations between the proportion of items keyed for socially desirable responses in the EPI and MMPI scales and the other dependent variables. No matter whether we look at the first factor loadings of the MMPI and EPI scales or the correlations of the MMPI and EPI scales with either of the two SD scales, the proportion of items keyed for

[4] The second factor loadings of MMPI scales correlated .67 with the corresponding second factor loadings of the EPI scales.

socially desirable responses in the scales is a fairly accurate predictor of the factor loadings and of the correlations of the scales with the SD scales.

The most parsimonious explanation of the correlations given in Table 1 would appear to be in terms of the fact that corresponding MMPI and EPI scales have exactly the same proportion of items keyed for socially desirable responses. When this proportion is large in an MMPI scale and, consequently, in an EPI scale, then high scorers on the SD scales also tend to obtain high scores on the EPI and MMPI scales and, as a result, the correlations of both the EPI and MMPI scales tend to be positively correlated with the SD scales. When the proportion of items keyed for socially desirable responses in an MMPI scale is small, then high scorers on the SD scales tend to obtain low scores on both the MMPI and EPI scales and the correlations of the MMPI and EPI scales with the SD scale both tend to be negative. The correlations given in Table 1 can thus be accounted for in terms of a common trait, the tendency to give socially desirable responses in self-description and by the fact that some MMPI and corresponding EPI scales are better measures of this trait than other MMPI and corresponding EPI scales by virtue of the imbalance in the social desirability keying of the items in the scales.

TABLE 2

Frequency Distribution of the Obtained Correlation Coefficients between 57 Paired MMPI and EPI Scales (f) and the Correlations Corrected for Attenuation (f′)

Correlations	f	f′
.90–.99		2
.80–.89		4
.70–.79	5	9
.60–.69	6	11
.50–.59	13	6
.40–.49	8	8
.30–.39	11	8
.20–.29	7	5
.10–.19	7	4

Table 2 shows the frequency distribution of the correlation coefficients between the paired MMPI and EPI scales. Test-retest reliability coefficients were available for each of the 57 MMPI scales

based upon responses of another sample of 95 male college students.[5] Test-retest reliability coefficients were not available for the 57 MMPI scales. However, assuming that the coefficients for the MMPI scales represent upper-bounds for the corresponding EPI scales, the correlations between the paired MMPI and EPI scales were corrected for attenuation using the reliability coefficients for the MMPI scales. The frequency distribution of the corrected correlations is given in the last column of Table 2.

In terms of social desirability considerations we are interested in accounting for the variation in the magnitude of the correlation coefficients between paired MMPI and EPI scales. That is, we would like to understand why some pairs of MMPI and EPI scales have relatively high correlations and other pairs have relatively low correlations. Table 3 shows the degree to which the correlations between the paired MMPI and EPI scales are related to the K-R 21 values of the MMPI and EPI scales, the absolute values of the first factor loadings of the MMPI and EPI scales, the absolute values of the correlations of the MMPI and EPI scales with the SD scales, and the product of the signed first factor loadings of the paired MMPI and EPI scales. Table 3 also gives the intercorrelations of these variables.

It has already been shown in Table 1 that the first factor loadings of MMPI and EPI scales are highly correlated. Thus, when an MMPI scale has a high positive loading on the first factor, so also does the corresponding EPI scale and the product of the factor loadings is also high and positive. Similarly, when an MMPI scale has a high negative loading on the first factor, so also does the corresponding EPI scale and the product of their first factor loadings is also high and positive. Table 3 shows that the products of the first factor loadings of corresponding MMPI and EPI scales correlate .85 with the observed correlations between the corresponding MMPI and EPI scales. Table 3 also shows that the higher the absolute value of the correlation of an MMPI scale with the SD scales, the higher the correlation of the MMPI scale with the corresponding EPI scale. In other words, when an MMPI scale has either a high positive or a high negative correlation with the SD scales it, in turn,

---

[5] I am indebted to Lewis R. Goldberg and Leonard G. Rorer of the Oregon Research Institute for calculating and making available to me the test-retest coefficients of the MMPI scales.

TABLE 3

Intercorrelations between the K-R 21 Values of 57 MMPI and 57 EPI Scales, the Absolute Values of the First Factor Loadings of the Scales, the Absolute Values of the Correlations of the Scales with Two SD Scales, the Product of the Signed First Factor Loadings of the Scales, and the Correlations between Corresponding MMPI and EPI Scales

	MMPI K-R 21	EPI K-R 21	MMPI I	EPI I′	MMPI SD_1	EPI SD_1	MMPI SD_2	EPI SD_2	I × I′	$r_{EPI\text{-}MMPI}$
MMPI K-R 21	—	.78	.66	.72	.65	.71	.62	.70	.75	.73
EPI K-R 21		—	.71	.74	.75	.74	.70	.72	.80	.82
MMPI I			—	.87	.96	.82	.95	.83	.94	.75
EPI I′				—	.84	.98	.82	.95	.95	.79
MMPI: SD_1					—	.82	.95	.84	.91	.71
EPI: SD_1						—	.78	.95	.92	.75
MMPI: SD_2							—	.84	.90	.74
EPI: SD_2								—	.91	.74
I × I′									—	.85

tends to have a relatively high positive correlation with its EPI counterpart.

It has already been shown that the first factor loadings of both MMPI and EPI scales are related to the proportion of items keyed for socially desirable responses in the scales. It has also been shown that MMPI scales which have high positive correlations with the SD scales are those which also have a large proportion of items keyed for socially desirable responses and that MMPI scales which have high negative correlations with the SD scales are those which also have a large proportion of items keyed for socially undesirable responses. Thus, the magnitude of the correlation between paired EPI and MMPI scales is related to the imbalance in the social desirability keying of the scales. MMPI scales which have either a large proportion of items keyed for socially desirable responses or a large proportion of items keyed for socially undesirable responses tend to have higher correlations with their EPI counterparts than MMPI scales which have a better balance in their social desirability keying.

Reasons have been presented elsewhere (Edwards, 1964; Edwards, Walsh, and Diers, 1963) to show why, if the tendency to give socially desirable responses is operating, personality scales in which the trait keyed responses are also consistently socially desirable or consistently socially undesirable responses should have a

higher degree of internal consistency than scales which are more balanced in their social desirability keying. The results reported in Table 3 are in accord with this prediction both with respect to the MMPI and EPI scales. Both MMPI and EPI scales which have high absolute correlations with either of the two SD scales tend to have a higher degree of internal consistency, as measured by the K-R 21 values of the scales, than scales which have low correlations with the SD scales. Since the important characteristic of corresponding MMPI and EPI scales which have high absolute correlations with the SD scales is that they have either a large proportion of items keyed for socially desirable responses or a large proportion of items keyed for socially undesirable responses, it follows that the internal consistency of the scales is related to the imbalance in the social desirability keying of the items in the scales. It may also be noted that the K-R 21 values of the EPI scales are related to the K-R 21 values of the corresponding MMPI scales, the correlation between the paired K-R 21 values being .78, and that, in general, the higher the K-R 21 value of an MMPI scale, the higher its correlation with the corresponding EPI scale.

The relationships found between the correlations of MMPI scales with the SD scale, the first factor loadings of the MMPI scales, the K-R 21 values of the MMPI scales, and the proportion of items keyed for socially desirable responses in the MMPI scales, in the present study, are in accord with the results of previous studies. Of greater significance, however, is the fact that essentially the same relationships between these variables are obtained when the scales involved are the EPI scales consisting of non-MMPI items.

In view of the fact that each of the items in the EPI was selected solely on the basis of the SDSV and P(T) of a corresponding item in the MMPI, it does not seem reasonable to believe that the items in each of the 57 EPI scales are similar in content to the content of the items in each of the corresponding 57 MMPI scales. If corresponding EPI and MMPI scales do not have a common content, they do have in common the same intensity of social desirability keying and the same proportion of items keyed for socially desirable responses and the latter variable has been shown to be a quite accurate predictor of the relationship between the correlation of MMPI and EPI scales with the SD scales and the first factor loadings of the MMPI and EPI scales.

Summary

An Experimental Personality Inventory (EPI) was developed in which each item was selected to match a corresponding item in the Minnesota Multiphasic Personality Inventory (MMPI) in terms of the social desirability scale value of the MMPI item and the probability of a True response to the MMPI item. The MMPI and EPI were given to 138 males and both inventories were scored for 57 MMPI scales. The first factor loadings and the correlations of the MMPI and EPI scales with two different SD scales were found to be highly correlated with the proportion of items keyed for socially desirable responses in the scales. The magnitude of the correlations between paired MMPI and EPI scales was found to be related to the imbalance in the social desirability keying of the items in the scales. MMPI scales with high Kuder-Richardson Formula 21 (K-R 21) values were found to have higher correlations with their EPI counterparts than MMPI scales with low K-R 21 values.

REFERENCES

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden Press, 1957.

Edwards, A. L. "Social Desirability or Acquiescence in the MMPI? A Case Study with the SD Scale." *Journal of Abnormal and Social Psychology,* LXIII (1961), 351–359.

Edwards, A. L. "A Factor Analysis of Experimental Social Desirability and Response Set Scales." *Journal of Applied Psychology,* XLVII (1963), 308–316.

Edwards, A. L. "Social Desirability and Performance on the MMPI." *Psychometrika,* XXIX (1964), 295–308.

Edwards, A. L. and Diers, Carol J. "Social Desirability and the Factorial Interpretation of the MMPI." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 501–509.

Edwards, A. L., Diers, Carol J., and Walker, J. N. "Response Sets and Factor Loadings on Sixty-One Personality Scales." *Journal of Applied Psychology,* XLVI (1962), 220–225.

Edwards, A. L. and Heathers, Louise B. "The First Factor of the MMPI: Social Desirability or Ego Strength?" *Journal of Consulting Psychology,* XXVI (1962), 99–100.

Edwards, A. L., Walsh, J. A., and Diers, Carol J. "The Relationship between Social Desirability and Internal Consistency of Personality Scales." *Journal of Applied Psychology,* XLVII (1963), 255–259.

Goldberg, L. R. and Rorer, L. G. "Test-Retest Item Statistics for Original and Reversed MMPI Items." *Oregon Research Institute Research Monograph* III (1963).

Messick, S. and Jackson, D. N. "Desirability Scale Values and Dispersions for MMPI Items." *Psychological Reports,* VIII (1961), 409–414.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

A FACTORIAL STUDY OF THE MMPI FOR STUDENTS IN HEALTH AND REHABILITATION[1]

HARRY E. ANDERSON, JR.
University of Georgia
HUGH C. DAVIS, JR. and WILLIAM D. WOLKING
College of Health Related Professions
University of Florida

Introduction

The growth of various health related professions has presented an attending problem in the guidance and selection of students planning to enter one of these professions, such as medical technology, occupational therapy, and physical therapy. The concentrated educational programs for these professions begin, for the most part, at the junior level in the university. There are, however, some rather general, introductory courses that cut across all of the professions and are required at the freshman and sophomore levels. It is becoming increasingly important to study characteristics and attributes of these beginning students for examining problems of differential program selection, dropouts, laboratory and clinical performance, and related aspects of students success.

The College of Health Related Professions (HRP) at the University of Florida represents a constellation of educational programs for the health related professions. A majority of the undergraduate students are females who, at the freshman and sophomore level, plan to enter occupational therapy, physical therapy, or medical technology. The present study is an investigation of the per-

[1] This research was supported in part by a research grant from the Vocational Rehabilitation Administration, Department of Health, Education, and Welfare, Washington, D. C.

sonality structure of these students for comparisons with other students and for use in future research.

Method

Sample

A sample of 168 female students was used in this study. All of the students were freshmen and sophomores in the HRP College. Moreover, all the students were enrolled in the basic course, "Introduction to Health Related Services," during the years 1962 to 1964.

Variables[2]

The ten basic clinical scales, excluding the four validity scales,[3] of the Minnesota Multiphasic Personality Inventory (MMPI) were used in the study. The ten scales are as follows: Hypochondriasis (*Hs*), Depression (*D*), Hysteria (*Hy*), Psychopathic Deviate (*Pd*), Masculinity-Feminity (*M-F*), Paranoia (*Pa*), Psychasthenia (*Pt*), Schizophrenia (*Sc*), Manic-Depressive (*Ma*), and Introversion (*Si*). The *Hs*, *Pd*, *Pt*, *Sc*, and *Ma* scales were used in K-corrected form, a usual scoring procedure in use of the MMPI and considered to have negligible effects on the interrelationships among the personality scales.

Analysis[4]

Scores for the 168 students on the 10 scales were intercorrelated by the product-moment method. The 10×10 correlation matrix, with squared multiple correlations in the principal diagonal, was factored by the principal component method, and the resulting factor structure was rotated by the normalized varimax routine (Kaiser, 1958).

The main interest in the study is the interrelationship among personality variables, and not necessarily the interrelationships

[2] Part of the data for this study was collected under NIMH Project Grant, 380, the Public Mental Health Methods in a University.

[3] Only seven cases exceeded a *T*-score above 70 on the K scale, and none appeared above 70 on the F and L scales.

[4] The analyses of data for this study were made possible through the use of personnel and facilities at the Computer Center of the University of Florida and the Computer Center of the University of Georgia.

among scoring artifacts in the MMPI scales. Items overlap among the scales, for instance, biases the data such that one could get a set of intercorrelations (e.g., Dahlstrom and Welsh, 1960, p. 82) among the scales if responses were made from a table of random numbers; we shall, therefore, refer to these as random correlations. Since the random correlations are not equal across all pairs of scales, there is no single base from which obtained correlations can be compared, and there is no obvious immediate way to estimate the impact of the random correlations on the underlying factor structure of the clinical battery. There are several indirect ways to estimate the impact, such as the one by Welsh (1952) wherein he removed some of the overlapping items in an attempt to factor somewhat more "pure" scales, but this method leaves very few items in some of the scales which tends to reduce standard deviations, reliability, and possible correlational relationships obtainable with the scales. Another method, which will be used herein, is to add or subtract, as the case may be, the random correlations from the obtained ones; squared correlations, however, will be used directly in the adding and subtracting process since they represent predictable variance estimates. The resulting "corrected" correlation matrix, with the original squared multiple correlations throughout the principal diagonal, was factored and factor-rotated by similar computer routines, and the resulting rotated factor structures were compared by the canonical correlation method. (e.g., Anderson, In Press; Kendall, 1957, Rao, 1952). This method allows for assessment of the impact of the random correlations on the underlying factor structure with consequent confidences in the evaluation of a personality structure obtainable with the MMPI.

Results

The obtained means, standard deviations, and inter-correlations, with squared multiple correlations in the principal diagonal, for the 10 MMPI scales are presented in Table 1. The means and standard deviations, presented in *T*-score units, look strikingly similar to the standardization norms for the MMPI and suggest confidence in the comparability of this sample to the standardization sample. It is the underlying factor structure, however, that will be of greatest interest.

For ease of comparisons, the corrected correlations are pre-

TABLE 1

*Intercorrelations, Means, and Standard Deviations of the MMPI Scales**

Scales	Hs	D	Hy	Pd	M-F	Pa	Pt	Sc	Ma	Si	Means	Standard Deviations
Hs	7219	5496	8154	4675	3303	4619	6206	5880	2504	3430	50.3155	9.73
D	5496	6808	4639	3674	1295	5868	7018	5362	0340	7141	50.2679	10.60
Hy	8154	4639	7302	5576	3275	5275	5813	5876	2790	2575	55.0357	9.653
Pd	4675	3674	5576	4883	2442	4507	5334	6224	4283	2076	55.0476	11.204
M-F	3303	1295	3275	2442	1883	1246	2419	3475	2759	0818	48.0655	10.154
Pa	4619	5868	5275	4507	1246	50797	6176	5854	2199	4971	52.9762	9.709
Pt	6206	7018	5813	5334	2419	6176	7434	7609	3350	6267	54.2738	9.629
Sc	5880	5362	5876	6224	3475	5854	7609	6954	4533	4635	53.8095	9.244
Ma	2504	0340	2790	4283	2759	2199	3350	4533	3390	0146	55.1667	11.815
Si	3430	7141	2575	2076	0818	4971	6267	4635	0146	5930	50.5179	11.301

* Decimals omitted from correlations.

sented in Table 2. The major reductions of correlations are between *Hy* and *Hs* from .82 to .67, for *Pt* and *D* from .70 to .65, between *Pt* and *Sc* from .76 to .71, and for *Ma* and *D*, from .03 to .18. This matrix also contains the original squared multiple correlations throughout the principal diagonal.

The correlation matrices in Tables 1 and 2 were factor analyzed separately by the principal component method. In accordance with the computer program cirterion, extracting principal axes with eigenvalues greater than zero, four factors, as shown in Table 3, were required to trace the matrix of uncorrected correlations in Table 1, while six were required for the corrected correlations. Actually, Factor IV in Table 3 and Factors IV through VI in Table 4 appear to be residual factors.

TABLE 2

*Corrected Intercorrelations of the MMPI Scales**

Scales	Hs	D	Hy	Pd	M-F	Pa	Pt	Sc	Ma	Si
Hs	7219	5193	6733	4671	3303	4609	6186	5825	2504	3424
D	5193	6808	4084	3397	1261	5854	6518	5268	0600	7084
Hy	6733	4084	7302	5277	3219	5213	5666	5751	2724	2728
Pd	4671	3397	5277	4883	2450	4132	5197	6015	4164	2086
M-F	3303	1261	3219	2450	1883	1179	2411	3463	2713	0795
Pa	4609	5854	5213	4132	1179	5079	6110	5383	2143	5007
Pt	6186	6518	5666	5197	2411	6110	7434	7114	3296	6152
Sc	5825	5268	5751	6015	3463	5383	7114	6954	4161	4608
Ma	2504	0600	2724	4164	2713	2143	3296	4161	3390	0717
Si	3424	7084	2728	2086	0795	5007	6152	4608	0717	5930

* Decimals omitted.

There is a general reduction in the communalities from the analysis in Table 3 to that in Table 4. As might be expected, the

TABLE 3

*Factor Matrix**

Scales	I	II	III	IV	Communalities
Hs	7777	1596	−3732	0812	7761
D	7381	−4463	−0736	0067	7494
Hy	7718	2705	−3441	−0800	7937
Pd	6469	2812	1456	−1442	5395
M-F	3326	2754	−0010	2153	2328
Pa	6993	−1476	0602	−1917	5511
Pt	8682	−1356	1381	0514	7938
Sc	8332	1269	2288	0448	7647
Ma	3747	4210	2973	0357	4073
Si	5930	−5389	0881	0912	6581

* All decimals omitted.

greatest loss in communality is about six to seven percent associated with the *Hy*, *Hs*, and *Ma* scales; the loss in *Sc*, *Pd*, and *Pt* is about three to five percent; and the communality changes in *D*, *M-F*, *Pa*, and *Si* are negligible. In most cases, however, the communality loss is minimal, and, in general, the first three factors in Table 3 resemble very closely the three factors in Table 4.

TABLE 4

*Factor Matrix**

(From Corrected Correlations)

Scales	I	II	III	IV	V	VI	Communalities
Hs	7576	1501	−3218	1648	−0748	−0010	7328
D	7267	−4616	−0414	0391	0015	−0197	7448
Hy	7395	2703	−3133	−1483	0682	0071	7447
Pd	6333	2987	1334	−1175	−0487	−0220	5248
M-F	3332	2850	−0026	1937	0742	−0009	2352
Pa	6933	−1609	0031	−2026	0028	0042	5476
Pt	8576	−1081	1092	0214	−0415	0314	7622
Sc	8186	1389	1860	0491	0256	−0140	7273
Ma	3795	3844	2953	0240	−0006	0096	3797
Si	6129	−5183	1008	0638	0395	0033	6600

* All decimals omitted.

Both sets of factors were machine-rotated by the varimax routine. The rotated factors in Table 5 are from the factor structure

in Table 3; those in Table 6, from the factor structure in Table 4. Again, the first three factors in Table 5 correspond very close to the three factors in Table 6; all other factors are again residual and therefore will be excluded from further analyses.

TABLE 5

*Rotated Factor Matrix**

Scales	Factors I	II	III	IV	Communalities
Hs	3570	2590	7573	0897	7761
D	8080	0430	3072	−0162	7494
Hy	2721	3362	7757	−0705	7937
Pd	2594	5766	3363	−1635	5395
M-F	0400	3429	2619	2123	2328
Pa	5833	2827	2906	−2157	5511
Pt	7135	4305	3147	0176	7938
Sc	5182	6310	3130	0114	7647
Ma	0078	6284	1104	0156	4073
Si	8060	0157	0696	0590	6581

* All decimals omitted.

TABLE 6

*Rotated Factor Matrix**
(From Corrected Correlations)

Scales	Factors I	II	III	IV	V	VI	Communalities
Hs	3681	2780	6802	1887	1473	0004	7328
D	8218	0564	2527	0360	0275	−0204	7448
Hy	2655	3416	7386	−0435	−1007	0000	7447
Pd	2379	5811	3460	−1009	0040	−0234	5248
M-F	0331	3411	2513	2338	0023	0003	2352
Pa	5847	2480	3307	−1746	−0662	−0005	5476
Pt	6815	4309	3287	0147	0533	0319	7622
Sc	4931	6090	3287	0712	−0033	−0140	7273
Ma	0323	6063	1004	0289	0004	0110	3797
Si	8068	0511	0607	0520	−0051	0033	6600

* All decimals omitted.

Interpreting factors is always a difficult task. The personal-social correlates given below will provide meaningful stereotypes to the extent that an individual's MMPI profile is weighted with the particular factor and within the limits of the validity of the MMPI research on non-clinic samples of college females (Drake and Oetting, 1959; Black, 1953).

Starting with the highest weights and moving to lowest, the first rotated factor is determined by *D, Si, Pt,* and *Pa. Sc* has a moderate loading on this factor too, but has slightly higher loading on Factor II. Young women with high loadings on this factor would tend to be seen by others as: shy, sensitive, sentimental, prone to worry, serious, natural, and with interests centered in home and family. Social submissiveness and emotional warmth in a serious prone-to-worry person summarizes the main traits. Under stress or in a college counseling situation, these women are likely to be insecure, self-conscious, and indecisive. They feel a lack of skill with the opposite sex and are especially distractible in their studies and tense on examinations.

The second rotated factor has major loadings on *Sc, Ma,* and *Pd* with *Pt* showing an appreciable loading also. It is of interest that the *M-F* scale has its highest loading on this factor, although the absolute value of its loading is not high. Black (1953) studied the ratings of normal college women with MMPI profiles weighted like this factor and found a mixture of flattering and unflattering terms used by their peers and themselves. Other girls saw these women as thoughtful, idealistic, and persevering, but also as self-centered and infantile, boastful and fickle, unemotional and self-dissatisfied. In self-descriptions, these women said they were polished, relaxed and thoughtful, but also said they were secretive, eccentric, and gloomy, and inarticulate. Compared with Factor I, this factor is defined by the characteristics extroversion, social assertiveness, independence of thought and action, and a tendency toward unsettled personal identification and philosophy. Under stress or mild breakdown conditions, these women will likely be confused, restless, verbal, and resistant in counseling or authority relationships. The loading on *Ma* alone seems to signify a relatively well socialized expression of energy. To the extent that the *Sc* loading is represented, the picture moves toward a more persistent and less socially desirable hypomania.

Determined almost exclusively by *Hy* and *Hs,* Factor III represents a combination of loadings which is relatively common in the MMPI profiles of many normal and clinic groups. Again Black's (1953) work with normal college girls suggests this factor will identify women with a general impunitive attitude in the way they view others, the world, and themselves. Their form of ex-

pression is often bland and innocuous. In sharp contrast, other college girls describe these people as selfish, self-centered, and having many physical complaints. They also labeled them as neurotic, dependent, indecisive, high-strung, hostile, irritable, and lacking in self-control. Thus, normal women with this pattern see themselves as cheerful, outgoing and optimistic—looking on the bright side of everything. Because of their repressive tendencies, their self-preception is often incomplete and inaccurate. Others see them as flighty, immature, and sometimes naively hostile people who may develop physical symptoms under stress. They are resistant to implications of psychological disorder. In the counseling situation, Drake (1959) has found these college women outgoing people, who are able to talk about their problems easily. They are often lacking in academic drive, find themselves blocking and tightening on examinations, and preoccupied with conflicts with their parents and home. Their lack of academic drive seems to be related to their rather strongly marriage-oriented view of college life. Because of this fact, it would be interesting to compare the academic success of this latter group with the success of students loading higher on the other two factors.

As noted previously, the two rotated factor structures, except for the residual factors, are quite similar in terms of variables with high and low, or moderate, loadings. A cursory examination would indicate, therefore, that the random correlations do not have much effect on the basic factor structure of the MMPI, but the method of canonical correlation is a much more systematic method for comparisons.

Factor scores were obtained for the 168 subjects on each of the rotated factor structures, using the first three factors, in Tables 5 and 6 by the formula, in matrix form

$$\mathrm{F} = \mathrm{ZS} \tag{1}$$

where, for p variables, N persons, and k factors, F is a $N \times k$ matrix of factor scores, S if the $p \times k$ matrix of factor loadings, and Z is the $N \times p$ matrix of standardized scores. Formula (1) is a modified form of Harman's (1960, p. 341) formula (16.10).

Harman's formula involves the inverse of the variables' correlation matrices and produces sets of orthogonal factor scores which would tend, operationally, to maximize similarities between corresponding factors in terms of the canonical correlation analysis.

Formula (1) above will produce correlated factor scores that differ from one set of factors to the other only in terms of S, the matrix of factor loadings. (For discussion of a similar point, see Pinnean and Newhouse, 1964, pp. 272–273.)

The factor scores' intercorrelations are presented in Table 7 which, for all intents and purposes herein, contains a supermatrix as shown by the partitioned blocks. The upper left and lower right matrices in Table 7 contain respectively the correlations

TABLE 7

Factor Score Correlations

FACTORS		U FACTORS			C FACTORS		
		I	II	III	I	II	III
U	I	1.0000	0.8475	0.9082	0.9999	0.8546	0.9103
	II	0.8475	1.0000	0.9350	0.8454	0.9998	0.9409
	III	0.9082	0.9350	1.0000	0.9071	0.9398	0.9997
C	I	0.9999	0.8454	0.9071	1.0000	0.8526	0.9091
	II	0.8546	0.9998	0.9398	0.8526	1.0000	0.9452
	III	0.9103	0.9409	0.9997	0.9091	0.9452	1.0000

among the U factors (from Table 5) and among the C factors (from Table 6); the lower left and upper right matrices, being transposed of each other, contain the cross-correlations between the two sets of factors. The correlations between the corresponding factors I, II, and III in U and C, respectively, have correlations all approaching unity in value.

The canonical correlations between the two sets of factors are presented in Table 8. Notwithstanding the high inter- as well as

TABLE 8

Canonical Correlations

Canonical Variables	Canonical Variables 1	2	3
1	0.9999	0.0004	−0.0001
2	0.0004	0.9996	−0.0001
3	−0.0001	−0.0001	0.9947

intra- correlations between individual pairs of factors, three very high, distinct correlations appear between the two sets of data. This result shows that there are three linear compounds, combining the

factor scores in each set of factors, that will produce extremely high correlations between the two factor structures, which further substantiates the fact that the random correlations have little, or no, effect on the internal structure of the MMPI for this study. Moreover, the occurrence of three extremely large, independent canonical vectors, as indicated by the near-zero off-diagonal correlations in Table 8, lends some credence to the system being three dimensional in structure, but a more detailed examination is required.

The equations for combining the U and C factors (from Table 7) to produce each of the canonical correlations are presented in Table 9; for instance, by entering each subject's factor score ap-

TABLE 9

Equations for Producing Canonical Correlations from Individual Factor Scores

Canonical Correlations	EQUATIONS
FIRST	$U_1 = 0.4292\ I + 0.3419\ II + 0.2670\ III$
	$C_1 = 0.4176\ I + 0.3446\ II + 0.2740\ III$
SECOND	$U_2 = -1.9715\ I + 1.5489\ II + 0.4263\ III$
	$C_2 = -1.9912\ I + 1.5686\ II + 0.4163\ III$
THIRD	$U_3 = -1.2791\ I - 2.3330\ II + 3.5417\ III$
	$C_3 = -1.2796\ I - 2.6137\ II + 3.8159\ III$

propriately into the equations, U_1 and C_1 will correlate at .9999, U_2 and C_2, at .9996, and U_3 and C_3 at .9947. The nearly identical weights for factors I, II, and III in U and C respectively occurs because the pairs of factors are each so highly related as to be co-linear, for all intents and purposes. The order of relative importance of the factors in determining the relationship between the two structures in the first canonical correlation, as indicated by their weights, is I, II, and III. The order is the same in the second canonical correlation, except that Factor I has an opposite (negative) sign. In the third canonical correlation the order is reversed, except that here the third factor has the opposite sign.

Notwithstanding the reluctance of previous researchers (e.g. Wheeler, Little, and Lehner, 1951) to account for scoring artifacts in the MMPI scales, it would appear from the results of the present study that such artifacts do not have a great deal of effect on the internal structure of the battery for the ten basic clinical scales.

If more scales are developed from the same set of items, however, the amount of item overlap might indeed become a problem and consequently, affect the internal factorial structure, such as in the study of Kassebaum, Couch, and Slater (1959). The method of canonical correlations, in any case, is a useful method for comparing resulting structures.

REFERENCES

Anderson, H. E., Jr. "Regression, Discriminant Analysis, and Standard Notation for Basic Statistics." Chapter 5 in Raymond B. Cattell's, *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally, (In Press).

Black, J. D. "The Interpretation of the MMPI Profiles of College Women." Unpublished thesis, University of Minnesota, 1953.

Dahlstrom, W. G. and Welsh, G. S. *An MMPI Handbook*. Minneapolis: The University of Minnesota Press, 1960.

Drake, L. E. and Oetting, E. R. *An MMPI Codebook for Counselors*. Minneapolis: University of Minnesota Press, 1959.

Harman, H. H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1960.

Kaiser, H. F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, XXIII (1958), 187–200.

Kassebaum, G., Couch, A., and Slater, P. The Factorial Dimensions of the MMPI." *Journal of Consulting Psychology*, XXIII (1959), 226–236.

Kendall, M. G. *A Course in Multivariate Analysis*. London: Griffin and Company, 1957.

Pinnean, S. R. and Newhouse, A. "Measures of Invariance and Comparability in Factor Analysis for Fixed Variables." *Psychometrika*, XXIX (1964), 271–281.

Rao, C. R. *Advanced Statistical Methods in Biometric Research*. New York: John Wiley and Sons, 1952.

Welsh, G. S. "A Factor Study of the MMPI Using Scales with Item Overlap Eliminated." *American Psychologist*, VII (1952), 341.

Wheeler, W. M., Little, K. B., and Lehner, G. F. "The Internal Structure of the MMPI." *Journal of Consulting Psychology*, XV (1951), 134–141.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

SIMPLEX STRUCTURE IN THE GRADING OF ESSAY TESTS[1]

ALBERT E. MYERS
Yale University
CAROLYN B. McCONVILLE and WILLIAM E. COFFMAN
Educational Testing Service

In December 1963, the 80,842 candidates electing the English Composition Test (ECT) in the College Board test administration were required to write a 20-minute essay. The test was the first ECT to include free writing since December 1956 and the first to include a 20-minute essay since April 1947 (CEEB, 1947). In a period of five days during the subsequent week, 145 readers gave each of the essays two independent readings. This study was designed to determine whether under these conditions the reading reliability would approach that obtained previously under somewhat less demanding circumstances. It culminates a 15-year period of research and development during which the College Board English Composition Test has moved from an all-essay examination through various stages of objective and semi-objective format to its present specifications.

The decision to reintroduce a 20-minute essay had not been made arbitrarily. Using data collected in 1961, Swineford (Noyes, 1963) had demonstrated that 20-minute essay scores based on five independent readings contributed uniquely to the validity of a one-hour English Composition Test for predicting a reliable essay criterion. The reading had been carried out during a five-day ex-

[1] We would like to thank Ledyard R Tucker for his immeasurable aid in the analysis of this experiment, Paul B. Diederich, Frances Swineford, and Karl G. Jöreskog for their comments on the manuscript, and Lynn K. Gaines for her assistance in gathering the data.

perimental reading by 25 carefully selected CEEB readers, and an analysis of the time required had indicated that as many as three readings might be obtained at a cost equivalent to that for reading the semi-objective interlinear exercises used periodically since 1951. The following year Godshalk, Swineford and Coffman (1965) demonstrated that Swineford's original findings could be duplicated during a half-day reading session by the 146 readers assembled to read the December 1962 interlinear exercise. They also found that substantially all of the unique validity of the 20-minute essay could be obtained from only two or three readings instead of four or five, particularly if a four-point scale were substituted for the original three-point scale.

The basic question asked in the current study was whether or not reader reliability would hold up over the five-day period required for reading over 80,000 essays. If so, presumably the validity findings of the experimental studies could be expected to apply. A related question of interest concerned the extent to which readers actually did follow the instructions to give a "single global judgment" rather than to respond to details of an essay. Subsequent sections of this report present the data which have led us to conclude that the operational reading was as reliable as previous experimental readings and that the readers were giving global judgments.

Method

Readers

Of the 145 readers, two were "chief readers" and 18 were designated as "table leaders." The chief readers were responsible for the conduct of the reading in general. The table leaders took care of any minor problems that arose or questions from the readers at their tables. Their main function, however, was to "spot-check" their readers and warn them when they seem to be grading consistently too high or too low. The remaining 125 readers read and graded the essays. The readers represented both public and private high schools and colleges. There were 42 women and 103 men.

Instructions

The general instructions to the readers were to read the essays

"holistically" and then assign a grade, 1 to 4. The interpretation of these grades was as follows:

1. obviously below a reasonable standard
2. not sufficient promise or competence to be considered in the upper half
3. clearly competent, promise of effective performance
4. superior; not perfect but very good; effective.

The essay was to be graded on its impact. Individual petty annoyances or pet peeves (e.g., spelling errors, split infinitives, etc.) were to be ignored as much as possible. Each table had its own letter code for the grades assigned, so that a second reader from a different table would not know the grade assigned by the previous reader.

Experimental Design

The experimental design called for a "control" group of papers to be read each day by a specific number of readers. Twenty-five papers were selected from the incoming essays representing as wide a geographical distribution as possible. The essays were hand-reproduced so that the end-product was indistinguishable from the originals. In this way the control papers could be presented to the readers along with the other essays without detection. A random selection of 25 judges per day was made prior to the conduct of the experiment. Each reader was presented a set of the 25 control papers, mixed with the other essays he was to read. Unforeseen circumstances (e.g., some judges' failure to read all of the control set of 25 papers on the day they were presented for reading) caused a variation in *N*s from 18 to 25 over the five-day period. The readers graded the controls just as they did the others, on a four-point scale, letter-coded their grade, entered their reader's number, and went on to the next essay.

Results

Reliability of the Judgments

The means and standard deviations of the 25 papers are shown for each day in Table 1. The variation of means was well within that expected by chance. An analysis of variance procedure (Winer, 1962) was used to assess the reliability of the judgments

TABLE 1

Means and Standard Deviations for the Five Reading Days

Day	Mean	Standard Deviation
1	2.07	.81
2	2.06	.70
3	2.05	.73
4	1.93	.73
5	2.14	.70

made by the subjects. This procedure provides an estimate of the average correlation between judges by comparing the variance between papers with the variance within papers.

TABLE 2

Summary Analyses of Variance for Each Day of the Reading

Day	Source of Variation	SS	df	MS	F	r[a]	n
1	Between papers	182.34	24	7.59	21.08	.466	23
	Within papers	199.74	550	.36			
2	Between papers	119.99	24	5.00	11.90	.364	19
	Within papers	186.84	450	.42			
3	Between papers	170.48	24	7.10	25.36	.493	25
	Within papers						
4	Between papers	165.18	24	6.88	23.72	.476	25
	Within papers	173.44	600	.29			
5	Between papers	86.78	24	3.62	9.28	.264	23
	Within papers	216.70	550	.39			

[a] These reliabiity coefficients have been stepped down by use of the Spearman-Brown formula. They represent the reliability of a single judgment.

Separate analyses were made on the data for each day. Table 2 provides a summary of the five analyses of variance which were used to determine the reliability of the readings for each day. The first step in each analysis was to calculate the reliability based upon the judgments of all the readers. By use of the Spearman-Brown formula these reliabilities were stepped down in order to find the reliability of single judgment. This single judgment coefficient represents the average reliability among all readers and across all papers. It can be seen that these reliabilities range from a high of .493 on day three to a low of .264 on day five. These coefficients, of course, represent the degree to which the readers could agree upon the quality of each paper; mean differences across days, of course, lower the overall reliability.

There is not, unfortunately, any well-developed procedure for determining the extent to which the variations among these five reliabilities can be attributed to chance alone. An appropriate test would require the ability to make multiple comparisons among F ratios. It is possible to evaluate the difference between two F ratios through a procedure presented by Schumann and Bradley (1957).

The F ratios shown in Table 2 were compared using the procedure and the tables presented by Schumann and Bradley. The reliability on day five was found to be significantly different from the reliabilities on days one, three, and four ($p < .05$).[2]

The .364 reliability found on day two was not significantly different from any of the other four reliabilities in the experiment. For this reason it is very difficult to determine whether the low reliability found on day five represents a significant change due to psychological factors or whether it, along with the reliability found on day two, was simply a low estimate of the population reliability. Given the large number of judgments that went into the estimation of these reliabilities, the most parsimonious conclusion would seem to be that day five does in fact represent a significant change from the reliabilities found in the other four days and that day two represents a low estimate of that four-day reliability.

TABLE 3

Reliabilities for Single and Multiple Readers

Day	Reliabilities				Coffman and Swineford Reliabilities for Four Readers	
	Single Reader	Two Readers	Three Readers	Four Readers	Topic A	Topic B
1	.466	.635	.723	.777		
2	.364	.533	.631	.695		
3	.493	.660	.744	.795		
4	.476	.644	.731	.784		
5	.264	.417	.518	.589		
Entire Reading	.406	.577	.672	.732	.71	.67

[2] Table presented by Schumann and Bradley is quite incomplete in that it does not provide for any level of significance other than .05. In addition, it does not provide for F ratios that have more than 30 degrees of freedom. It was necessary in the present analysis, therefore, to interpolate well beyond those values given in the table. Consequently, the only significance level which will be used in those analyses involving the Schumann and Bradley procedure will be .05.

Table 3 shows how these reliabilities can be stepped up with multiple readings. If two readers read each paper, the reliability jumps from .406 to .577. Since five days were required to obtain two readings, perhaps special attention should be paid to the two-reader reliability. The four-reader reliabilities were about the same as those found by Godshalk, Swineford and Coffman (1965) (also shown in Table 3) and demonstrated that the extended reading period did not lower the reliabilities. With the exception of day five, in fact, the present coefficients tended to be higher.

Factor Analysis

Even though the readers in this study have been instructed to make a single global judgment, there was, of course, no guarantee that they were not operating on different bases. In their study of essay gradings Diederich, French, and Carlton (1961) had concluded that the readers had made their global judgments on the basis of diverse "schools of thought." It is obvious that an overall reliability coefficient would be reduced if the readers in a sample were using different points of view in making their judgments.

A factor analysis of the *covariances* among the 25 papers was done to assess the possibility that the 118 readers in the present study have a small number of identifiable points of view (similar to Tucker and Coffman, 1959). Four factors were extracted from the 25×25 matrix and a varimax rotation performed. The results of this analysis are shown in Table 4. The reader should note that the communalities and loadings are not based on unit variances. Rather, each paper has a variance which is related to the mean rating received by the paper. Since these values are less than unity, the communalities appear small for those accustomed to using communalities in factor analyses of correlation matrices. The unrotated data are given in the appendix.

With a few exceptions, quite good simple structure was achieved with the varimax rotation. That is, a paper was loaded fairly high on one factor and reasonably low on all the others. When the papers were grouped according to the factors they seemed to represent, it was noted that there were great similarities in the average grade received by those papers. In Table 5 it can be seen that the poor papers tended to load on factor two while the good papers tended to load on factor three.

TABLE 4

Varimax Factors

Paper	Communalities	SD of Papers	Factor 1	Factor 2	Factor 3	Factor 4
1	18	61	−01	33	08	27
2	12	59	00	08	12	31
3	07	62	−01	13	24	04
4	13	68	−03	18	−01	31
5	08	52	02	27	09	−04
6	13	54	−01	15	26	20
7	12	60	14	−02	25	18
8	11	57	06	10	30	−04
9	13	54	08	08	−01	34
10	30	68	10	−04	54	00
11	11	52	04	27	13	14
12	10	48	06	29	08	06
13	06	61	07	−05	21	09
14	14	60	08	34	04	14
15	14	66	21	06	12	27
16	11	58	16	25	−01	15
17	13	54	11	31	−05	13
18	06	48	23	10	03	01
19	15	55	38	−03	09	05
20	22	59	44	12	−01	11
21	24	65	36	33	−06	−01
22	19	56	22	36	−11	04
23	11	59	25	00	07	21
24	22	67	44	11	11	04
25	12	51	32	07	09	−02

The temptation to start naming factors at this point was successfully resisted. Instead, papers were grouped according to the factors they loaded on. Analyses of variance were computed for the group of papers identified with each factor for each day. In other words, the same type of analysis as was done earlier was repeated for the four factors. Two papers, Numbers 21 and 22, were not included in these analyses because they had such high loadings on two of the factors. Since there were four factors and five days this meant, of course, that 20 such analyses were performed.

The immediate interest in these analyses was to investigate the discriminability that existed between papers that loaded on the same factor. The median F ratio (out of the total of five, one for each day) is shown for each factor in Table 6. Although all these values were significant at at least the .05 level, there were considerable differences in the degree to which the clusters of papers were discriminable from each other. The ω^2s which indicate the percent

TABLE 5

Means of Papers on the Various Factors

Factor	Paper	Mean Score	Range of Mean Scores
I	18	1.96	
	1[illegible]	2.12	
	20	1.93	
	23	1.91	.21
	24	2.08	
	25	2.09	
II	1	1.69	
	5	1.87	
	11	1.77	
	12	1.82	.62
	14	1.46	
	16	1.52	
	17	1.25	
III	3	2.58	
	6	1.99	
	7	2.46	
	8	2.42	1.35
	10	3.08	
	13	3.34	
IV	2	2.12	
	4	2.04	
	9	1.79	.66
	15	2.45	

TABLE 6

Analyses of Variance to Assess Homogeneity of Means of Papers on Each Factor

Factor	Source of Variation	SS	df	MS	F	ω^2
1	Between Judges	15.10	22	.69	2.52	.05
	Between Papers	2.06	5	.41		
	Residual	17.94	110	.16		
	Total	35.10	137			
2	Between Judges	16.66	18	.92	3.66	.10
	Between Papers	4.30	6	.72		
	Residual	21.13	108	.20		
	Total	42.09	132			
3	Between Judges	8.49	18	.47	18.90	.42
	Between Papers	23.13	5	4.63		
	Residual	22.04	90	.24		
	Total	53.66	113			
4	Between Judges	16.54	24	.69	6.61	.11
	Between Papers	4.59	3	1.53		
	Residual	16.66	72	.23		
	Total	37.79	99			

of variance which was determined by the discriminability between papers (Hays, 1963) were quite different for the four factors. Factor three, which included all of the good papers, had 42 percent of the variance of the judgments determined by the discriminability between papers. The next largest ω^2 was .11 for factor four. As far as significance testing is concerned, the F ratios for papers defining factor three were significantly larger than the ratios for any of the other factors and the F ratio for factor four was significantly larger than that for factor one.

Discussion

When the factor analysis had been completed, the papers were grouped by factors and read by the authors and others. There certainly did not seem to be any obvious reason why the papers were grouped this way, except for their general quality. Poor papers tended to be poor on every attribute of composition and good papers tended to be good on these attributes. Students who have difficulty organizing papers also have difficulty in spelling, etc. It seemed that the factors did not represent content dimensions and that some other type of explanation for the factor structure would be necessary.

In contrast to what is usually done in a factor analytic study, no attempt will be made here to provide labels which will identify the factors. On the contrary, we will present an argument to the effect that these factors do not represent different sources of content. We will argue that they represent diverse points of view in the relating of grades to the quality of a paper.

Let us consider how we might have interpreted the data if only factors 2 and 3 had emerged. That is, suppose all the good papers were clustered together and all the poor papers were clustered together in a two-factor space. It would not matter in this situation how any individual assigned grades, just so long as he tended to give all the good papers the same type of score and to give all the bad papers the same type of score. Thus, some readers might give a score of 1 to the poorer half of the papers; others might assign a 2 to this half; and there might even be a few who give a 3 to all of the poorer papers. Similarly, they would assign the same score to the better half (2, 3, or 4). Suppose further that all the readers had perfect agreement as to which were the good papers and which the

bad. Figure 1 shows the hypothetical function that would describe scoring of this type. In this situation there is no discriminability among papers in the good group nor is there any discriminability among papers in the poor group. A factor analysis of these data would show that there were two factors, and those factors would undoubtedly be identified as good papers and poor papers.

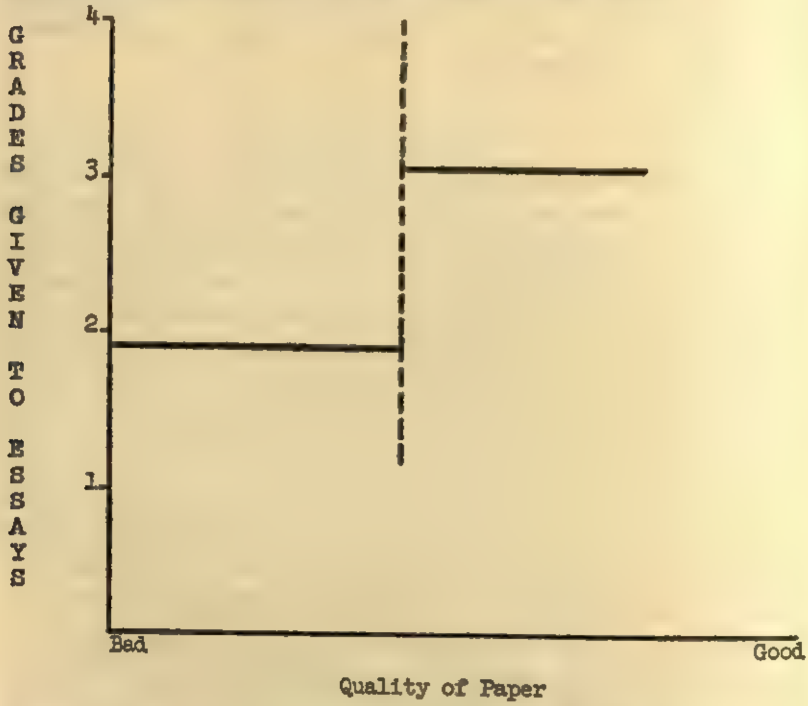


Figure 1. Hypothetical relationship between quality of paper and grade assignments with no differences in discrimination among papers at a given level.

Similarly, if there were three categories of papers instead of two, and if all the judges continued to be equally discriminating within the three categories, we would find that the data would produce three factors.

A fourth factor could be introduced if there were differences among the readers with respect to the location of the range between the two cutoff points. Thus some papers that were in the middle range for some readers would be in either the upper or lower categories for other readers, not because they valued the papers differently, but because they simply used different cutoff

points for classification. The shifting of the middle range up and down would cause certain papers to correlate which were quite disparate in mean score while they maintained correlations with papers with similar scores. This is precisely what occurred with those papers which loaded on factor 4.

It should also be noted that there were differences in the discriminability among the papers found on a factor. There was a good deal of discriminability among the better papers and very little among the poorer ones. This indicates that the function relating the quality of the paper to the grade assignment was curvilinear.

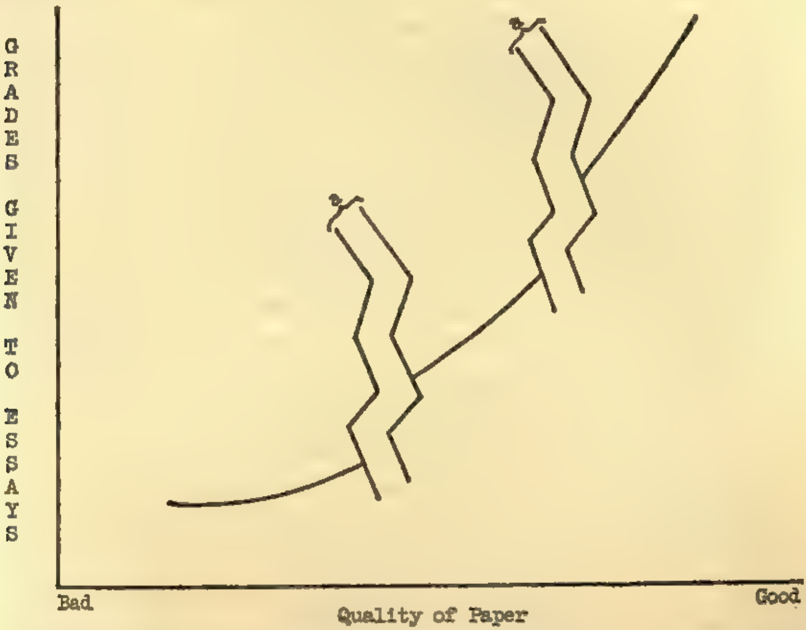


Figure 2. Hypothetical relationship between quality of paper and grade assignments with variable discrimination.

[a] The reader will note that the range incorporated here represents the discrepancy between the most lenient and the most severe reader. The most lenient reader includes many papers in the highest category while the most severe reader includes few; the opposite is true with papers in the low category.

Figure 2 depicts a functional relationship between grades and quality which is very descriptive of the data found in this study. The breaks at the cutoff points are intended to communicate that

there is a cutoff range which is the composite of all the readers used in the study. It can be seen that there is differential discriminability among papers in various parts of the range. Data of this sort would produce four common factors.

Explaining the factor structure in this manner does not imply that there is not a variety of attributes used in the judgment. The judges, to be sure, must have been sensitive to differences in grammar, spelling, general style, etc. These are the attributes which the judges used as criteria to make their assignment to the categories good, bad, and mediocre. Why then, we might ask, aren't these attributes represented by the factors instead of the function relating quality and grade assignment? Why have we chosen to represent these data as if they described a simplex?

Guttman (1954) has described a simplex as a structural relationship that exists between elements when these elements are ordered on a single dimension of complexity. Complexity is used in this context in a most general way; a complex element is made up of many factors while a less complex element is made up of fewer factors. In the present setting, essays which have good spelling, good grammar, good ideas, good style are complex in the sense that they manifest the existence of many positive attributes. A poor paper, on the other hand, is "simple" in that none of these attributes may be present. Thus, the ordering of papers by their mean could be construed as an ordering by complexity. It has been pointed out already, of course, that with the present data an ordering by factors is almost equivalent to ordering by means.

A description of the data in these terms is most satisfying in that it suggests that the readers were making global judgments with respect to quality using a set of interrelated criteria. This is consistent with their instructions and the fact that they took about a minute to read each paper. It also suggests, in contrast to the Diederich, French, and Carlton study, that, although all the readers may have assigned different weights to the various attributes of a composition (e.g., spelling, grammar, organization), each of the individual attributes tended to be given the same weight by the readers. That is, it does not seem as though some judges rated the papers primarily upon grammar, while other judges rated the papers primarily upon style. Rather, this explanation suggests that the relative weighting between grammar and style is *roughly* the same for all judges.

The Problem of the Fifth Day

The reliabilities found in the present study were slightly higher than those found in previous investigations of the short essay. The severe drop on the fifth day, however, does pose certain practical problems. Although it is quite understandable that the readers should suffer a loss in vigilance as they neared the end of this fairly grueling task, our understanding, per se, does very little to combat the problem. The drop in reliability to .26 on the last day poses a serious problem. The immediate question becomes, therefore, Is it possible to maintain a high state of vigilance at the end of the reading period? If the readers are mature and conscientious people, as they were in the present setting, they might resolve the problem themselves if it were pointed out to them that there was a tendency for them to slip at the end of the reading period. If, however, the readers were not able to maintain a state of vigilance by their own efforts, then the reliability problem will appear more troublesome.

We have been assuming that the drop in reliability occurs at the end of a reading period because the readers are anticipating the completion of their task. If this is so, it means that there would be an equivalent drop in reliability regardless of how long the reading period was. This implies immediately, of course, that this problem cannot be handled by simply shortening the reading period by any small amount. It would seem that some external source would be needed to bolster the reader morale and effort. Future research efforts should be concerned with experimental evaluation of such external efforts.

REFERENCES

College Entrance Examination Board. Forty-Seventh Annual Report of the Director. New York: Author, 1947. Pp. 23–24.

Diederich, P. B., French, J. W., and Carlton, Sydell T. "Factors in the Judgments of Writing Ability." Research Bulletin 61-15. Princeton: Educational Testing Service, 1961. (Multilithed Report)

Godshalk, Fred, Swineford, Frances, and Coffman, W. E. *The Measurement of Writing Ability*, Research Monograph No. 6. New York: College Entrance Examination Board, 1966.

Guttman, L. "A New Approach to Factor Analysis: the Radex. In

Lazarsfeld, P. F. (Ed.), *Mathematical Thinking in the Social Sciences*. Glencoe, Illinois: Free Press, 1954.
Hays, W. L. *Statistics for Psychologists*. New York: Holt, Rinehart and Winston, 1963.
Noyes, E. S. "Essay and Objective Tests in English." *College Board Review*, Winter 1963, No. 49.
Schumann, D. E. W. and Bradley, R. A. "The Comparison of the Sensitivities of Similar Experiments: Theory. *Annals of Mathematics and Statistics*, XXVIII (1957), 902–920.
Tucker, L. R. and Coffman, W. E. A Factor Analytic Study of Judged Relevance of Test Items. Research Memorandum 59–11. Princeton: Educational Testing Service, 1959. (Multilithed Report)
Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

EFFECTS OF COACHING ON THE COLLEGE BOARD ENGLISH COMPOSITION TEST

PAUL I. JACOBS
Educational Testing Service

If college selection is to be based in part upon achievement test scores, it is important that such scores have these characteristics:

1. The achievement test scores of students should be increased by taking the course of study for which the test has been developed.
2. At least some of the increase should be relatively permanent.
3. Any procedure that affects achievement test scores should similarly affect the college criterion measure.

One set of procedures for which there is some question regarding all three points is that of "coaching" practices. The everyday usage of the term "coaching" generally refers to intensive teaching, of either individuals or groups of students, that takes place outside of the regular school program of instruction, and that is specifically directed toward obtaining high test scores, rather than toward a good education.

The present study provides information regarding the effects of coaching on the College Board English Composition Test (ECT).

Method

Design

The basic experimental paradigm was to contrast performance on the English Composition Test of a group of students given coaching with that of an otherwise comparable group of students not given coaching. Students were randomly assigned to the two groups. Scores on the Verbal section of the Preliminary Scholastic Aptitude Test (PSAT-V) were already available for the students,

and served as an additional check on the initial comparability of the groups. The paradigm, then, was that of a Pretest-Posttest Control Group Design (Campbell and Stanley, 1963). This basic paradigm was complicated in the following ways:

1. To assess how possible effects of coaching might vary with the situation, comparable coached and non-coached (control) groups were set up within each of six different secondary schools. In this way the composite effects of differences among the schools in procedures, personnel, and students could be examined in a nonanalytic way, that is, without isolating specific sources of variability between the schools in coaching effects.

2. One might reasonably expect within a given school "leakage" of coaching, that is, the coached students passing on the benefits of their extra instruction to their non-coached fellow students. Such leakage would reduce the estimated effects of the coaching. To guard against this possibility, each of the six schools containing both coached and control students was matched with another "control-only" school, the matching based upon type of school (public, parochial; city, suburban), and mean SAT scores of the schools' College Board candidates in a preceding year (January and March, 1958). The pairing of coached-and-non-coached schools with control-only schools, then constituted a Nonequivalent Control Group Design (Campbell and Stanley, 1963).

3. One might also reasonably expect that certain item types on the ECT would be more coachable than others. For this reason an attempt was made to estimate the relative coachability of the three item types that appeared on a given form of the ECT. Each coaching group received coaching on only two of the three item types upon which they were later to be tested. Item types were randomly assigned to schools, with the restriction that each of the three possible pairs of item types be assigned to two of the six schools. In this way the interaction of school with item type was deliberately confounded (Lindquist, 1953).

4. To estimate the permanence of possible coaching effects, students were given a second ECT ten months after the first.

Materials

The ECT given in May 1961 was used as the basic criterion measure. It contained three item types: *sentence correction* (SC),

construction shift (CS), and *paragraph organization* (PO). The ECT given in March 1962 served as the second criterion measure. It also contained *sentence correction* and *paragraph organization* item types, as well as *error recognition* (ER) items. Examples of all these item types are given in Jacobs (1964).

The coaching materials provided to the cooperating schools were actual items taken from retired (no longer secure) forms of the ECT. The construction shift items are publicly available (Thomas, 1956). A total of 45 sentence correction, 16 construction shift, and 16 paragraph organization paragraphs (comprising 103 separate paragraph organization items) were used as coaching materials.

Procedure

Selection of Schools. It was desired to work with schools within the northeast metropolitan area, where the problem created for students, parents, and school officials by the existence of coaching schools is probably greatest. From a list of schools in the New York City-New Jersey-Philadelphia area that had 100 or more Scholastic Aptitude Test (SAT) candidates in January and March 1958 and that gave the Preliminary Scholastic Aptitude Test to its juniors, six pairs of schools were selected. Each pair was roughly matched on mean SAT Verbal scores of their 1958 candidates. A letter was sent to the principal of each of these schools, explaining the purpose and general design of the study. He was asked to consent to cooperate without being told first whether his school would be selected to have both coached and control groups or just a control group. It was felt that if the principal were given this information before he agreed to cooperate, there might be a lower rate of cooperation among the control-only schools approached, which might in turn produce some systematic initial differences between the coached-and-control schools and the control-only schools. It turned out that when the letters were followed up with telephone calls, all the principals were willing to cooperate, and expressed mild to intense enthusiasm. In the making of further arrangements with the schools, one school was discovered not to be able to cooperate because a free classroom for coaching was unavailable during the school day lengthened by multiple sessions, and a thirteenth school was contacted and accepted.

A brief description of the cooperating schools is given in Table 1.

TABLE 1

A Description of Schools Cooperating in the Study

Code Name	Type and Location	Mean SAT-V of Candidates in Earlier Year	Mean SAT-M of Candidates in Earlier Year	Item Types Coached P.O.	C.S.	S.C.
A_1	Public H.S., Long Island, N. Y.	501	516	X	—	X
A_2	Public H.S., Long Island, N. Y.	501	529	—	—	—
B_1	Public H.S., New Jersey	467	495	X	X	X
B_2	Public H.S., New Jersey	479	507	—	—	—
C_1	Public H.S., New York City	486	502	X	X	—
C_2	Public H.S., New York City	486	495	—	—	—
D_1	Catholic H.S., New York City	507	536	—	X	X
D_2	Catholic H.S., Philadelphia	508	548	—	—	—
E_1	Public H.S., Westchester, N. Y.	531	532	X	X	—
E_2	Public H.S., Westchester, N. Y.	532	544	—	—	—
F_1	Public H.S., New York City	444	469	X	—	X
F_2	Public H.S., New York City	445	478	—	—	—

Selection of Coaches

In each cooperating coached-and-control school the chairman of the English Department selected a member of his department to serve as the coaching instructor, or coach. Each coach was given an honorarium of $5.00 an hour.

Selection of Students

Within each cooperating school an announcement was distributed to each junior thought by the school to be a likely college applicant. The student was asked to volunteer for the experiment,

knowing that through random selection he might not be selected at all for the experiment, and if selected, he might be placed in the control rather than the coached group. The student was further told that those selected for either the coached or control groups could take the ECT that year without paying the $6.00 fee, that they might be asked to take the ECT again the following year, also without fee, and that for those selected for the coached group there would be no charge for the coaching.

It was intended to select randomly from among the volunteers in each control-only school 30 students, and in each coached-and-control school, 60 students, to be further randomly assigned to the coached and control groups. In one coached-and-control school, however, an administrative mix-up resulted in there being altogether only 24 volunteers. All 24 were therefore accepted into the experiment, and 12 of them were selected at random for the coached group.

The Coaching

Each coach was presented with coaching materials for two item types. He was told that these were item types that had appeared on the ECT in the past and were likely to occur again. The coach was encouraged to supplement these materials with whatever additional material seemed appropriate, to give homework assignments, and to keep a log describing the coaching activity.

From these logs it appeared that in each school roughly one-half of the coaching time was devoted to direct work with the coaching materials supplied, while the rest of the time was devoted to English composition work not as specifically aimed at these item types. The coach used a wide range of supplementary material, including "cram" books, the booklet *A Description of the College Board Achievement Tests,* and Strunk and White (1959).

The coaching was given three hours a week for six weeks. In some schools student motivation lagged, resulting in spotty attendance at the coaching sessions. In School B_1, on the other hand, student interest and speed of working was such that the coach requested additional coaching materials from ETS, and was given the exercises dealing with the third item type. In another school (School C_1) both coached and control students were routinely

given a "Composition Conference Period" that presumably overlapped in function with the coaching sessions.

After the sixth week of coaching, all students in the study took the ECT. In the spring following the first ECT, the students in the coached-and-control schools, now seniors, were again contacted to take the ECT. As will be seen, the results available at that time suggested it would be pointless to retest the students in the control-only schools.

Results and Discussion

The sample of schools, located as they were in the New York City-New Jersey-Philadelphia metropolitan complex, and having a large number of College Board candidates, were well chosen for the purpose of the study. One might argue, however, that the students within these schools who volunteered to take part in the experiment as juniors were not part of the usual College Board candidate population, and so the results obtained with them might not be generalizable to this population.

If the "right" students volunteered to take the ECT free of charge, we would expect these same students to pay for and take the SAT as seniors. A check of the ETS records was made, and the results are presented in Table 2. This supports the assumption that the students volunteering to take part in the experiment were members of the usual College Board candidate population.

TABLE 2

Number of Students in Experiment Taking SAT between December 1961 and August 1962 in Each Coaching-and-Control School

School	Number of Students
A_1	54
B_1	54
C_1	44
D_1	57
E_1	22*
F_1	35

* In each of these schools except E_1 the total number of students in the experiment was 60; in E_1 it was 24.

The next analysis was to estimate the possible effects of leakage of coaching from coached to control students within the

same school. Leakage would be demonstrated by the control group in a coached-and-control school doing better than the control group matched with it in a control-only school.

Control group means of ECT scores in the two types of schools are shown in Table 3. The differences in means were relatively small, and not consistent in direction. We conclude that they reflect imperfect matching of schools and error of measurement. Since leakage appeared to be negligible, the students in the control-only schools were not asked to retake the ECT as seniors. It was assumed that any leakage would already have occurred by the time the first ECT was taken.

TABLE 3

Comparison of Control Group Means on ECT in Each Type of School

School	Coaching and Control School	Control School Only
A	475	498
B	495	487
C	457	455
D	482	496
E	554	551
F	418	422

It was appropriate, therefore, to compare the coached with the control students within each coached-and-control school to estimate the possible effects of coaching. An initial analysis made was of total test score on the three item types of the ECT. The results are presented in Table 4. The scores are given in terms of the College Board scaled score (mean $= 500$, $\sigma = 100$). In two of the six coached-and-control schools, there was essentially no difference between coached and control groups, while in each of the other four schools, the mean difference of from 44 to 73 points favoring the coached group was statistically significant and also presumably practically significant. We have, then, rather clear evidence of over-all coaching effects in some schools, and not in others.

The next analysis was aimed at exploring the mechanism by which the coached groups achieved their superiority. Did they get fewer items wrong and/or omit fewer items, and was this confined to those item types for which coaching was received? Two statistically independent part scores were derived from each item type

TABLE 4

Total Test Score Results on ECT_1

School	PSAT-V-Mean Coached Group	PSAT-V-Mean Control Group	ECT_1 Mean Coached Group	ECT_1 Mean Control Group	Analysis of Covariance	*df*	Adjusted M.S.	*F*
A_1	47.8 (29)***	46.5 (25)	519.3	474.8	Between	1	16378.1	4.24*
					Within	51	3864.9	
B_1	43.8 (24)	43.1 (27)	540.9	495.1	Between	1	20444.0	4.70*
					Within	48	4348.1	
C_1	47.4 (11)	44.1 (24)	456.4	457.2	Between	1	3510.6	—
					Within	32	5454.9	
D_1	44.4 (25)	46.1 (22)	475.4	481.8	Between	1	526.7	—
					Within	44	3532.9	
E_1	55.2 (12)	55.6 (12)	627.0	554.3	Between	1	33601.0	10.75**
					Within	21	3126.5	
F_1	44.0 (26)	41.3 (23)	480.0	418.3	Between	1	20013.3	6.96*
					Within	46	2877.5	

* $p < .05$
** $p < .01$
*** Number of students.

for each group: the number of items answered incorrectly, and the number of items not attempted. The results are shown in Table 5.

Ignoring item types, it appears that in each school showing an over-all coaching effect, the coached group achieved its superiority by answering fewer items incorrectly; regarding omits, there is no tendency common to each of these four schools. A tendency for coaching to increase the number of omits on the SAT has been noted previously (French, 1955). It is interesting that in School C_1, which does not show an over-all coaching effect, the coached group shows the same pattern of fewer wrongs as seen in those schools that do show an over-all coaching effect, and shows almost double the number of omits.

Does this pattern occur only on those item types for which there was coaching? This seems to be true only regarding number *wrongs* for School F_1.

The general conclusions, then, from this inquiry into the mechanism of coaching effects, are:

1. Coached groups achieve their superiority by getting fewer items wrong.
2. In general this effect is not limited to those item types coached for.

TABLE 5

Number Wrongs and Number Omits by Item Type for Each Group

School	Item Type	Wrongs Coached	Wrongs Control	Omits Coached	Omits Control
A_1	S.C.	11.3	14.1	5.2	3.4
	P.O.	12.1	15.3	4.2	4.1
	C.S.*	9.0	12.0	6.8	5.9
	Total	32.4	41.4	16.2	13.4
B_1	S.C.	8.4	13.0	6.2	5.6
	P.O.	11.6	12.6	4.2	4.6
	C.S.	7.5	8.4	8.6	8.9
	Total	27.5	34.0	19.0	19.1
C_1	S.C.*	12.4	15.0	8.9	3.4
	P.O.	12.2	18.5	9.0	2.2
	C.S.	9.8	11.2	8.4	7.7
	Total	34.4	44.7	26.3	13.3
D_1	S.C.	12.7	13.3	4.8	4.8
	P.O.*	15.0	14.3	6.0	4.5
	C.S.	10.0	11.2	6.4	6.4
	Total	37.7	38.8	17.2	15.7
E_1	S.C.*	9.4	11.5	1.4	3.2
	P.O.	7.8	11.1	0.5	1.7
	C.S.	6.5	10.5	4.5	3.8
	Total	23.7	33.1	6.4	8.7
F_1	S.C.	14.6	16.3	5.8	6.5
	P.O.	14.4	16.3	4.2	6.2
	C.S.*	12.5	11.8	6.2	7.0
	Total	41.5	44.4	16.2	19.7

* Not coached for at this school.

3. In one school not showing an over-all coaching effect there is evidence from the part score analyses of some effects of coaching.

Permanence of Coaching Effects

How permanent are the gains achieved by the coached group? A comparison of over-all scores on the retest of the ECT after ten months is shown in Table 6.

In Schools B_1, D_1, E_1, and F_1, both coached and control groups appeared to have gained as a result of practice and growth. The

TABLE 6

Total Test Score Results on ECT_2

School	PSAT-V Mean Coached Group	PSAT-V Mean Control Group	ECT_2 Mean Coached Group	ECT_2 Mean Control Group	Analysis of Covariance	*df*	Adjusted M.S.	*F*
A_1	46.4 (9)*	52.0 (5)	519.8	569.6				
B_1	43.9 (23)	44.8 (21)	570.8	550.8	Between	1	9258.0	2.32
					Within	41	3987.5	
C_1	41.5 (2)	32.5 (2)	428.5	437.5				
D_1	47.8 (17)	45.7 (15)	536.2	509.5	Between	1	159.4	—
					Within	29	3565.6	—
E_1	54.3 (11)	56.1 (8)	639.7	647.4	Between	1	539.7	—
					Within	16	5491.8	
F_1	49.5 (11)	42.8 (12)	557.1	491.7	Between	1	455.2	—
					Within	20	7297.5	

* Number of students.

covariance analyses, however, indicate that the gain for the control groups is large enough for them to catch up with the coached groups. Such differences between coached and control groups in mean score as are observed on ECT_2 are accountable for in terms of initial differences in PSAT-V scores. In Schools A_1 and C_1 the small number of cases do not permit any conclusions to be drawn. Although the pattern of results for each school differs rather widely, an overview of results on PSAT, ECT, and ECT_2 for all schools combined is shown in Figure 1.

In the present study intensive instruction involving in part the use of actual items from former examinations with motivated students from the College Board candidate population yielded an increment in scores relative to a noncoached control group. This increment had disappeared by ten months as measured by retest, and presumably would not affect the college criterion measure.

Does, then, the candidate who receives coaching have an advantage over the candidate who does not? The present results indicate that he may indeed, but the considerable variability from school to school in coaching effects suggests that this will depend on some specifics of the coaching situation. Perhaps any analysis of the social importance of coaching effects on an achievement test should compare these effects with the effects of ordinary teaching. The test-retest data of the control groups show an interest-

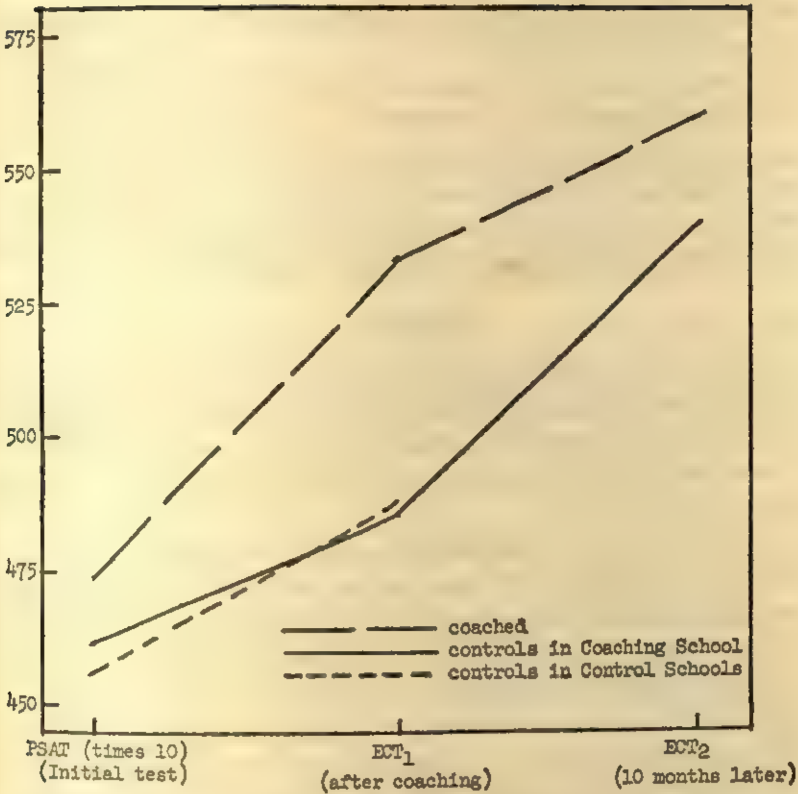


Figure 1. An overview of test results for all schools combined.

ing parallel here: ECT_2 minus ECT_1, which may be taken as an estimate for those students taking both tests of the combined effects of practice and of ten months' growth, varied from 15 points at School D_1 to 94 points at School E_1. Apparently the advantage accruing from an additional year's study of English will also depend on some specifics of the teaching situation. It may be, then, that an appropriate program of self-study by the individual would provide an increment equivalent to that obtained from coaching. The present study does not provide evidence on this point.

Curr and Gourlay (1960) have proposed an alternate definition of coaching, not in terms of a procedure, as in the present study, but in terms of observed effects:

(a) "Coaching is present when gains vary considerably with the type of test used to measure the basic skill to be acquired. Such variation in gains we would regard as evidence more of a

variation in the acquisition of test-skills than of a genuine improvement in the basic skill."

(b) "A gain can be attributed to coaching when it is transitory, i.e., when it disappears in a fairly short period of time—12 months or less." The gains in the present situation meet the second of these criteria, but as they were in general not confined to specific item types, they do not meet the first. In terms of Curr and Gourlay's definition, therefore, *no* coaching effects were found.

Summary

Coaching may increase students' scores on an achievement test without increasing, or increasing only temporarily, their skills in the area being tested. This study deals with the immediate and longer term effects of coaching on the CEEB English Composition Test.

Students at each of six cooperating high schools were randomly assigned to coached and non-coached conditions. Following the coaching sessions, all students took the English Composition Test.

The immediate effects varied from essentially no mean difference in test scores between coached and non-coached students at one school to a mean difference of 73 points favoring the coached students at another school. In general the coaching increment was not confined to the specific item types coached for. The coached groups achieved their superiority by getting fewer items wrong.

When coached and non-coached students were retested ten months later it was found that the coaching increment had disappeared. Estimates of growth and practice effects from the test-retest data of the non-coached students varied from a mean of 15 points at one school to a mean of 94 points at another.

It was concluded that the candidate who receives coaching may indeed have an advantage over the candidate who does not. But the variability from school to school both of coaching effects and of score gains due to "growth" indicates that the specifics of the teaching situation are quite important.

REFERENCES

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research on Teaching. In N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally & Co., 1963. Pp. 171–246.

Curr, W. and Gourlay, N. "The Effects of Practice on Performance in Scholastic Tests." *British Journal of Educational Psychology*, XXX (1960), 155–167.

French, J. W. *The Coachability of the SAT in Public Schools.* Research Bulletin 55–26. Princeton, New Jersey: Educational Testing Service, 1955.

Jacobs, P. I. *Effects of Coaching on the College Board English Composition Test.* Research Bulletin 64–24. Princeton, New Jersey: Educational Testing Service, 1964.

Lindquist, E. F. *Design and Analysis of Experiments in Psychology and Education.* New York: Houghton Mifflin Company, 1953.

Strunk, W., Jr. and White, E. B. *The Elements of Style.* New York: MacMillan, 1959.

Thomas, M. "Construction Shift Exercises in Objective Form." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVI (1956), 181–186.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

SOME HYPOTHESES CONCERNING TEST ADEQUACY[1]

DONALD W. FISKE[2]
University of Chicago

An earlier paper (Fiske, 1963) presented the rationale for using the cumulative homogeneity model in developing psychological tests, especially in the personality domain. It also outlined in non-technical form a proposed technique for assessing the adequacy of a test in terms of such a model.

This paper describes the analytical technique more specifically. It also summarizes the findings from application of the technique to various scales from a wide variety of existing tests. From these findings are derived generalizations in the form of hypotheses. These state the characteristics of a test which are believed to contribute to high degrees of adequacy.

The concern underlying this paper is measurement in the service of theory. The goal is to improve personality measurement so that personality concepts can be better studied and refined. The technical approach utilized for this purpose is unsophisticated. It involves simply a breakdown of the total observed variance in a set of test responses to determine the relative contribution of the means for persons and of the means for items, and also the remaining variance from all other sources, including especially the interaction between persons and items. A simple descriptive approach, rather than one using variance estimates, has been employed to

[1] This investigation was supported by a PHS research grant, MH 06582, from the National Institute of Mental Health, Public Health Service.
[2] I am indebted to Naomi Berne, Jere Brophy, Pamela Pearson, Castellano Turner, and Thomas Tyler for their suggestions and assistance. For making available some of the data used in this paper, I am indebted to Charles Dicken, Norman Endler, Gerald Goodman, J. McV. Hunt, Hugh Lane, and Anita Sandke.

emphasize the particular observations actually in hand and the particular measuring instrument. While the testing of a theoretical proposition requires inference from the sample measured to some loosely described population, we consider it dangerous to assume that any single personality test adequately represents the total domain of a personality construct. Stated in different words, most current personality constructs are broad and heterogeneous; following the strategy based on the cumulative homogeneity model, any one test should be designed and used to assess only one specified portion of such a construct. Appropriate combinations of such restricted sets of measuring operations can then be formed to represent a total construct, or a major segment of it. In principle, this procedure should make possible a close fit between a group of measuring operations and a construct. Maximizing the congruence between construct and operations seems essential to the empirical study of each construct, for the purpose of furthering the science of personality.

Assessing the Homogeneity of a Test

Below is an informal statement of computational procedures and indices for assessing the homogeneity of persons, the homogeneity of items, and the proportion of unwanted variance in the responses of a sample to a set of test items. The underlying rationale is developed in "Homogeneity and Variation in Measuring Personality" (Fiske, 1963).

The ideal is taken to be the following:

1. Optimizing the variance of person means (this goal will usually be equivalent to maximizing this variance since existing tests never approach the postulated ideal of a rectangular distribution).

2. Optimizing the variance of item means (again, the postulated ideal of a rectangular distribution requires much larger variance than is ever found).

3. Minimizing unwanted variance associated with instability and person-item interaction.

The variance of person means is used rather than the more usual variance of total scores for persons because the former is independent of n, the number of items. Similarly, the variance of item means is independent of N, the number of persons. For any test, the variance of person means can be compared to that for

item means. Also, the variances of person means (or item means) from several tests can be compared. Thus the adequacy of a test is evaluated in terms of the quality of the measurement obtained from the response of the typical person to the typical item. Once this quality has been judged as adequate, a decision about the suitable length of the test can be made in terms of such considerations as the desired level of the conventional index of internal consistency and of the amount of testing time available.

The Method

The analysis essentially consists of writing a data matrix, in which the rows are persons and the columns are items and each entry is the response of a person to an item, and of computing various sums of squares. These values are then used to obtain four coefficients and to break down the total variance into three parts.

The total sum of squares and the sums of squares for rows, for columns, and for remainder are computed in the standard fashion for a two-way analysis of variance with one entry per cell. Dividing each of the latter three partial sums of squares by its degrees of freedom yields the usual variance estimates, which provide a convenient way of obtaining certain coefficients.

The standard coefficient of internal consistency, r_{tt} (the general coefficient alpha [Cronbach, 1951]; for dichotomous scoring, Kuder-Richardson 20) is obtained by taking the complement of the variance estimate for remainder divided by the variance estimate for rows. The dual of this coefficient, r_{gg}, is obtained by taking the complement of the ratio, variance estimate for remainder over that for columns (cf. Guilford, 1954, pp. 383–385).

More useful than these coefficients are the corresponding ones for items and persons. The coefficient r_{tt} can be stepped *down* by the Spearman-Brown formula to obtain the intraclass correlation for items (which can informally be seen as the average correlation between pairs of items). Similar treatment of r_{gg} yields r_{pp}, the intraclass correlation between persons (or the average intercorrelation between pairs of persons).

In this approach, the variance indices are descriptive; they are stated in terms of actual variance in the sample (rather than as estimates of variance in the population). The total variance, s_t^2,

is the variance of the total distribution of entries of the data matrix. If the scoring is dichotomous (if only 0 and 1 are used as entries), the total variance is the proportion of 0's times the proportion of 1's (i.e., pq). If the scoring is not dichotomous, the total variance is obtained by dividing the total sum of squares by nN, the total number of observations. (This total variance should not be confused with the variance of person scores which is used in the standard formula for Kuder-Richardson 20.)

By dividing the sum of squares for columns by nN, we obtain s_{Mi}^2, the actual or obtained variance of the item means. Similarly, still dividing by nN, we obtain the actual variance of person means (s_{Mp}^2) from the sum of squares for rows, and the remainder variance (s_e^2) from the remainder sum of squares.

While the total variance for dichotomously scored tests usually is close to .25, it may be somewhat less if the group mean departs considerably from .5, the middle of the possible range of means. Moreover, it is desirable to be able to compare tests with dichotomous scoring and tests with multi-step scoring, where the total variance may be incommensurate with that from 0,1 scoring. Therefore the three component variances are stated as proportions of the total variance.

It should be noted that for a test with dichotomous scoring, it is not necessary to write out, or even to have available, the data matrix. The necessary computations can be performed from a distribution of person scores and a distribution of item endorsement values. The total variance is obtained from the grand mean as indicated above.

The computations described above provide an extensive set of indices for assessing a test in terms of the cumulative homogeneity model. The description of these operations is also intended to show certain mathematical relationships among the indices. If in actual practice, one is interested only in certain indices, some of these computations may not be necessary.

Indices of Test Adequacy

The most important single index is the proportion of remainder variance, s_e^2/s_t^2. This value indicates the total relative contribution of what is often called "error." We use the term "remainder" rather than "error" because we can specify some of the elements

included in this value. By subtracting the contributions of persons and of items from the total variance, we have a value which is the sum of the contributions from several unwanted sources. While these cannot be separately assessed with the data from a single trial, something is known of their approximate size from data with repeated trials.

In dichotomously scored tests, one part of the remainder variance is an artifact of such scoring. For any corresponding arrays of item and of person means, the expected value for each cell can be computed. The artifactual component is the mean squared deviation of such values from 0 or 1, whichever is closer. Thus the artifact stems from the restriction of the scoring to these two values (to a discrete scale) when a completely homogeneous data matrix requires that the cell entries be on a continuous scale. In the idealized case with rectangular distributions of person and of item means, this artifactual component reaches .33 as its relative contribution. It approaches .00 as the variance of either rows or of columns approaches zero. The more steps that are used on the scoring of each item, the less the contribution of this artifact.

Another large component in the remainder variance is person-item interaction. This source reflects the extent to which persons give responses to items which are different from those expected on the basis of the corresponding row and column means. In this analysis, we have already identified part of this average discrepancy as artifactual. The rest of this discrepancy can be attributed to true person-item interaction.

Conceptually, such interaction has two parts. One part may be attributed to variability of responses over time, or instability. Analyses of some data from repeated trials suggest that the proportional contribution from instability is relatively small. I.e., the departure of the response of a person to an item from his mean response over a series of trials contributes relatively little to the total variance. Such departures may be either toward the value predicted from row and column means or away from it.

The second and, conceptually, the most important part of person-item interaction is idiosyncrasy, the tendency of the person's response (or average response over trials) to be different from that predicted. This source is heterogeneity, the complement of what we seek in maximizing the homogeneity of persons and of items.

Such idiosyncrasy may be explained in various ways. It may stem from a highly unusual, individual interpretation of an item. It may also come from the variable, as defined by the test items, taking different forms in different persons. E.g., while, for the group as a whole, Item A may be endorsed more frequently than Item B, there may be a minority of individuals whose response tendency of the Item A variety is much weaker than their tendency along the lines of Item B.

The other two important pieces of information obtained from this analysis are reflected in the proportions of total variance contributed by the variances of item means and of person means. With dichotomous scoring, each of these values is .33 in what is postulated to be the ideal case. The relative variance of persons tells how well people are differentiated. The relative variance of items indicates the extent to which the items tap different amounts of the underlying variable, rather than clustering together near the average amount of the variable. Appreciable variation in item means is necessary to determine the comparability of response patterns for different persons.

In practice, it seems desirable for these two values to be of about the same magnitude. The approach taken in this paper requires that neither one be increased at the expense of the other. It is obvious that a test with considerable item variance is of little value if it does not differentiate among persons. On the other side, it is asserted that differentiation among persons should not be emphasized at the cost of item variance.

If the relative variance of item means is low, it can be increased by adding items with endorsement values as different as possible from those of existing items, to make the overall distribution more closely approximate a rectangular one. If the relative variance of person means is low, the test is not differentiating well for the given group of persons to whom it was administered. It may be that the test is inappropriate for the group's range of response strengths for the variable being measured. Or it may be simply a poor test.

The same essential information is, of course, contained in the actual variances of item and of person means.

If one prefers to study correlation coefficients rather than variances, r_{ii} and r_{pp} can be examined. With dichotomous scoring, each

of these two coefficients is .50 in the ideal case. r_{ii} is closely related to the variance of person means and r_{pp} bears a parallel relationship to the variance of item means. When $n=N$ (when the number of items equals the number of persons), the relation between the magnitudes of the variances for person and item means ($s_{M_p}{}^2$ and $s_{M_i}{}^2$, respectively) is reflected in the relation between r_{ii} and r_{pp}.

As an index of the usual notion of internal consistency, the consistency of items, r_{ii} is preferred to r_{tt} because the latter varies with the number of items. r_{tt} is an evaluation of the given set of items as put together in the particular test. r_{ii} indicates the adequacy of items of this kind for the group of persons tested.

r_{pp} indicates the extent to which the several people are producing the same pattern of responses, the pattern summarized in the ordering of the item means. In the ideal case, each individual's pattern will be consistent with that of every other individual and with that of the item means. Each person's pattern will have a general elevation determined by his mean score. This score and the scoring system may produce ties between items with different group means but will not produce reversals. Thus, for dichotomous scoring, each person's pattern will match that for each other person, except for a run of items, the length of which is equal to the difference between their total scores.

r_{gg} is of negligible value since, for any test, it varies with the number of subjects contributing to the data matrix.

(For a somewhat similar approach to the evaluation of a test, see Cronbach, Rajaratnam, and Gleser, 1963. The major differences are these: Cronbach and his associates are working from a relative frequency model rather than a cumulative homogeneity model; they use variance estimates rather than straight descriptive or obtained variances.)

Result and Interpretations

The Data

Table 1 summarizes most of the applications which have been made of these procedures for analyzing homogeneity. Where two or more scales from one instrument have been studied, the scales with the highest and the lowest proportional remainder variance ($s_e{}^2/s_t{}^2$) are usually presented. The various instruments are listed

TABLE 1

Homogeneity Indices for Various Scales

Test and Scale		N	$\bar{X}_{ip}$*	r_{ii}	r_{pp}	$\frac{s_{Mi}^2}{s_t^2}$	$\frac{s_{Mp}^2}{s_t^2}$	$\frac{s_e^2}{s_t^2}$
Edwards Personal Preference Schedule	n Deference	30	.36	.03	.07	.09	.06	.85
	n Achievement	30	.63	.02	.08	.10	.05	.84
Social Desirability (Edwards)		36	.93	.05	.08	.10	.07	.83
Kuder Preference Record	Literature	50	.47	.05	.14	.14	.06	.80
	Mechanical	50	.31	.07	.17	.17	.07	.76
California Psychological Inventory	Achievement via Independence	34	.76	.05	.09	.10	.06	.84
	Good Impression	34	.38	.09	.27	.27	.08	.65
Allport-Vernon-Lindzey	Theoretical (1–4)	20	2.86	.07	.12	.15	.10	.75
	Political (1–4)	27	2.29	.04	.19	.20	.07	.72
Thurstone Temperament Schedule	Emotionally Stable	30	.46	.03	.11	.13	.07	.80
	Sociable	40	.67	.11	.27	.26	.11	.63
Personality Research Inventory (ETS)	Attitude toward work	32	.53	.10	.05	.07	.18	.75
	Talkative	32	.40	.24	.12	.10	.28	.61
WAIS	Comprehension (0–2)	48	1.68	.06	.11	.11	.11	.78
	Similarities (0–2)	48	1.51	.13	.34	.30	.13	.56
Manifest Anxiety Scale		30	.40	.21	.16	.15	.19	.66
MMPI	Hy (Pre-therapy)	30	.42	.01	.20	.23	.02	.75
	Hy (Peace Corps)	36	.37	.01	.56	.57	.01	.42
Holtzman Ink-Blot	Hostility (0–2)	20	.29	.03	.07	.10	.12	.78
	Human (0–2)	20	.84	.03	.35	.35	.08	.57
Conation Indicator	Science	50	.57	.17	.20	.17	.24	.60
	Daring	50	.52	.17	.34	.28	.20	.52
Street-Gestalt		50	.60	.10	.31	.30	.08	.62
Physiognomic Cues (0–5)		50	1.80	.22	.38	.33	.16	.51
Weitzenhoffer-Hilgard Hypnotic Susceptibility		124	.44	.29	.23	.17	.29	.54
ACE	Verbal Completion	50	.69	.07	.32	.32	.07	.61
	Figure Analogies	50	.66	.07	.57	.55	.04	.41
Self-Disclosure Questionnaire	to Close Male Friend (0–2)	16	1.44	.26	.23	.23	.23	.54
	to Casual Acquaintance (0–2)	39	.86	.40	.44	.32	.28	.40
FIRO	Expressed Inclusion	76	.51	.25	.40	.32	.23	.45
	Wanted Control	76	.50	.28	.54	.43	.20	.37
S-R Inventory of Anxiousness	Total Test (1–5)	71	2.05	.12	.40	.38	.08	.54
Mode 8—(Not) Enjoy the Challenge (1–5)		71	2.68	.15	.12	.11	.20	.69
Mode 1—Heart beats faster (1–5)		71	2.70	.31	.43	.33	.25	.42
Situation 7—Sailboat on rough sea (1–5)		71	1.91	.28	.24	.19	.27	.55
Situation 4—Ledge on mountain (1–5)		71	2.58	.37	.42	.30	.29	.41
Situation 4—Four items for Distress factor (1–5)		71	3.18	.59	.37	.16	.58	.26

* Means greater than unity can occur for tests with multi-step (rather than dichotomous) scoring of items. Where such scoring was used, the score range is given after the name of the scale.

in descending order of this expression for the remainder variance (using the average value when two scales from the same instrument are presented). Omitted are various data for repeated trials and data from a few instruments for which only one scale was analyzed.

The remainder variance is stressed in this table and in the ensuing discussion because it is believed to be the best single index of the adequacy of a test. While differentiation among persons is highly desirable, minimal person-item interaction is essential for a good instrument designed to measure individual differences.

Data for two groups were available in some instances. When the two groups were highly comparable, the indices were typically very similar. When the conditions of testing or other factors made the groups not comparable, the corresponding indices differed to a greater or lesser extent. The largest observed differences, for the MMPI *Hy* scale, are given in the table.

The ACE, the Street-Gestalt, and the Physiognomic Cues were given with time limits. However, the great majority of the subjects attempted almost all of the items in these scales.

The data selected for Table 1 may yield an erroneous impression that there is a strong association between r_{ii} and r_{pp}, and also between s_{Mi}^2 and s_{Mp}^2. For unselected tests, such associations over tests are much smaller.

Inspection of Table 1 discloses a great range in each of the several indices: not only does the proportion of remainder variance vary from .85 to .26, but the proportions of variance associated with item and with person means vary from .09 to .57 and from .04 to .58 respectively. The smallest and largest values of the coefficients are .01 and .59 for items and .05 and .57 for persons. The problem is to infer the sources of these large variabilities.

The Degree of Structure in the Task Given to the Subject

One pertinent condition appears to be the degree of structure in the task which is set for the subject. There seems to be some negative relationship between such degree of structuring and the contribution of remainder variance.

In intelligence tests and in many cognitive-perceptual tests, the subject has a clear instruction about what he is to do and about the criteria he is to use in selecting his response. The standard per-

sonality questionnaires, on the other hand, provide a task with very little structure (cf. Fiske and Butler, 1963). The subject has two difficulties. First, he has vague criteria for determining what his response should be: for example, how often is "often"? Second, the process by which he arrives at his response, before coding it into the available alternatives, is not indicated explicitly. For instance, the question might be: do you find it difficult to give a speech before a large group? The subject can go about the process of answering it in several ways. He can compare this statement to his general picture of himself. He can recall one or several pertinent experiences and base his response on his memory of his feelings. He can also decide that an affirmative answer would be true for most people and therefore is true for him. (These are some of the processes discerned in reports from subjects responding to such an item.) Insofar as various subjects go about answering such a question in different ways, the task is not structured.

Another way of interpreting this condition is to consider whether the subject's task is almost wholly to react to present stimuli or requires the subject to scan his memories for pertinent instances of past behavior or feelings. Intelligence and cognitive-preceptual tests involve immediate processes and reactions while most questionnaires require mediation of the response process through prior experiences and memories. A seeming exception to the generalization that such a distinction is related to adequacy of the instrument is the S-R Inventory (Endler, Hunt, and Rosenstein, 1962). While its quality involves other considerations, the merit of this instrument may also stem in part from the fact that, for many of the presented situations, the subject must imagine himself in them since they are ones which few people have actually experienced.

When a task is structured for each subject, it is more likely to be approached in the same way by the several subjects. The greater the homogeneity or similarity between the subjects' approaches to a test, the more adequate the test in terms of the criteria used in this paper. This proposition is, at present, an hypothesis which has not been systematically tested.

For a given subject, the structuring of a task can be increased by familiarity with it. A person who has experienced a given test has had the opportunity to develop attitudes toward it. Such acquisition of stable perceptions and reactions to a test probably accounts

for the typically slight gains in adequacy of the various indices on second and later trials. (Only limited investigation of this effect has been carried out.)

A special case of structuring is found in the Hypnotic Susceptibility Scale (Weitzenhoffer and Hilgard, 1959). The several items in this scale are used to estimate the degree of trance present in the subject over the several minutes required for application of the test stimuli. This unique instrument highlights the way in which structuring contributes to homogeneity: in a structured task, the subject tends to approach the several items in the same way. The critical factor is this maintenance of set across items.

From these considerations, we derive the following hypothesis: *1. The adequacy of a test is a direct function of the degree of structuring of the task, structuring which is established for the subject by the instructions and the situation.*

We have not defined "structure" yet. It is a concept used here to refer to a common element in the several aspects of the testing situation considered above. Structuring imposes constraints on the subject so that, in the ideal case, the responses of each subject are determined solely by the degree to which he possesses the disposition being measured. To delineate the term more sharply and without circularity, we must consider it in two ways. A task is structured when, from the subject's point of view, he feels he knows (a) what the potential consequences of the testing experience are for him, what effect (if any) his performance in the situation will have upon him and (b) what he is expected to do. This latter expectation includes not only the way he is to approach each item-stimulus and the transaction between the item and him but also the criteria he is to use in selecting his response. From the examiner's viewpoint, a task is structured when the several subjects agree on the nature of the potential consequences of the testing and on these expectations about the task for the subjects, i.e., when the subjects perceive the same demand characteristics (Orne, 1962) in the situation. Hypothesis 1 refers to both subjective and objective forms of structuring.

As so defined, the degree of structure in a testing situation cannot be directly assessed by any one set of operations. Significant indirect evidence can be obtained from the consistency of the responses over items and persons and valuable information can

also be sought through carefully planned questioning after the test is completed.

The classification of immediacy versus mediacy, mentioned earlier, overlaps greatly with a classification proposed by Campbell (1957): the objective versus voluntary nature of the criteria to be used by the subject in selecting his response. Both classifications can be subsumed under the concept of structure. While tests with objective criteria are more structured, the large range of structuring among tests with voluntary criteria indicates the considerable contribution of other conditions such as immediacy-mediacy and those discussed below.

The Degree of Structure in the Item-Stimulus

Another condition contributing to test adequacy appears to be the degree of structuring present in the items. Such structuring can be present in two places: the question or stimulus to which the subject reacts and the response alternatives afforded him. The latter will be considered in the next section.

The stimulus in an intelligence test can be said to be highly structured. In this instance, structure is defined as univocality or lack of ambiguity, and can be assessed empirically. Items in intellegence and perceptual-cognitive tests have the same meaning to those who can give the correct answer. In contrast, the typical item in personality questionnaires is subject to diverse interpretations (Benton, 1935; Eisenberg, 1941; Loehlin, 1961). However, items in personality instruments can be highly structured by increasing the degree of specification. Thus the Jourard Self-Disclosure Questionnaire (Jourard and Lasakow, 1958) specifies not only the content of the possible disclosure but also the person to whom it is disclosed. The S-R Inventory asks for a report on the extent of each response tendency in each specified situation. To be sure, a composite score from such tests is affected by considerable interaction between persons and objects or between persons and situations. However, such an instrument has the potential advantages of providing specific subscores for objects or situations and of enabling one to estimate the extent of such interaction.

Lumping together explicit specification and lack of ambiguity in meaning, we formulate this hypothesis: *2. The adequacy of a test is a direct function of the degree of structuring of the typical*

item-stimulus. Structuring from the examiner's point of view can be determined in terms of the degree of consensus among the subjects in their perceptions or interpretations of the items, as reported under a separate set of instructions. To assess the extent to which the subject sees the item as structured, it is probably sufficient simply to ask him whether he felt any uncertainty about what the item meant. The first, objective type of structuring is necessary for test adequacy. The second, subjective form is probably desirable although it is not sufficient: even though subjects may not experience any uncertainty about an item, they may interpret it in different ways.

The data for the S-R Inventory reported in Table 1 illustrate a trend which may turn out to have very great significance. When the responses for each specified situation are analyzed as separate "tests," r_{ii} and s_{Mp}^2 (and also s_{Mp}^2/s_t^2) typically are considerably higher than the corresponding values for the total test. Preliminary data on an experimental test which also specifies situations show a similar picture. When each type of response is analyzed separately, the same indices tend to be more favorable but the other set of indices (r_{pp} and s_{Mi}^2) do not systematically improve. (In the S-R analyses for separate Modes of Response, the median values of these latter indices are substantially lower than those for the test as a whole.)

The interpretation of these comparisons remains to be established. It may be that, in such questionnaires, the delineation of a specific situation enables the subject to keep a single clear mental image in mind as he selects his responses to the several items. It may also be that behavioral dispositions are associated with (or determined by) the particular situation to a much greater degree than is recognized by most conceptual formulations. Whatever the explanation, the data strongly suggest that the variance of persons can be increased, and at least questionnaires can be made more adequate, by incorporating in groups of items a single explicit set of situational conditions.

The Degree of Structure in the Response Alternatives

One might consider the possibility that the adequacy of a test is in part a function of the degree of structuring of the responses which the subject is permitted to make. Here again, structure

means univocality. However, there is little evidence to justify viewing this aspect of a test as a separate property related to adequacy. It seems pertinent, for example, that the S-R Inventory must be viewed very favorably in terms of the criteria being used in this paper in spite of the fact that the subject is given a set of five ordered steps from which to select his responses, with only the first and last step defined at all, and these definitions being very general for the high ends.

More structured response alternatives are ordinarily found when the task is structured, as in intelligence tests, and when the items are structured, as in perceptual-cognitive tests. They also are more likely to occur when the criteria for selection of responses are objective rather than voluntary.

The Degree of Substantive Homogeneity

Underlying the methodological approach presented in this paper is the proposition that the adequacy of a measuring procedure used in research on a construct should be defined in terms of the substantive homogeneity of the test over items and over people. The substantive homogeneity of a test refers to the degree to which the test scores reflect a unitary conceptual entity. It is maximal when the same function or response tendency is measured by the several items in each of the subjects. The several criteria are proposed as indices of such homogeneity. The two hypotheses above state potential conditions contributing to this homogeneity: the structuring of the task elicits the same function in the several subjects as they cope with each of the items; the structuring or specificity of the items makes it possible to utilize items pertaining to a common response tendency.

Separate attention is being given to this basic concept because the item content is of overwhelming importance. The data in Table 1 suggest conditions contributing to substantive homogeneity. Among the more adequate tests are those in which a priori identification of content was given major emphasis. (This association was first observed by Lee J. Cronbach.) The S-R Inventory was formulated with no use of the techniques of test development and refinement. Psychometric properties played a secondary role in the construction of FIRO-B (Schutz, 1958). In contrast, empirical criteria for item selection were the sole or primary consider-

ations in the development of such tests as the Kuder Preference Record, the California Psychological Inventory, and the MMPI.

Factor analysis undoubtedly can be used to improve homogeneity (Comrey, 1961). The most favorable total pattern of indices found to date is that for one situation in the S-R Inventory and for four modes of response with high loadings on the "distress" factor emerging in the work of Levin (1963). On the other hand, factor analysis was used in the development of the Thurstone Temperament Schedule, an instrument which is only moderately successful in terms of our criteria.

Part of Table 1 suggests that competition between different variables in the same item is undesirable. In the EPPS, the two statements in each pair represent different needs. Each item in the first part of the Allport-Vernon-Lindzey requires the choice between two values; in the second part, four values must be ordered.

Notwithstanding the definitional aspect discussed above, the primacy of substantive homogeneity justifies the formulation of this proposition as a hypothesis: *3. The adequacy of a test is a direct function of the substantive homogeneity of its items.*

Some readers will argue that the maximizing of substantive homogeneity leads to the creation of scales which assess a very limited conceptual domain. Just as attitude scales with high homogeneity and reproducibility by Guttman's criteria tend to be narrow and redundant, so highly homogeneous tests may appear so specific that what they measure may seem trivial. Thus in several of the FIRO scales, the same stem is typically used for two items which differ only by presenting alternative sets of response choices (one set in terms of temporal frequency and the other in terms of the number of people with whom the subject manifests the tendency stated in the stem). Similarly, the S-R Inventory has its best sets of indices when some one situation or single mode of response is analyzed separately.

It seems probable that high homogeneity, as assessed by the proposed criteria, will typically be associated with a limited diversity of item content. Such an association, however, need not be an obstacle. The best strategy for optimal measurement of a personality variable involves several steps. First, the variable must be analyzed conceptually to determine the various forms and the various situations or contexts in which the behavioral tendencies ap-

pear. Then scales must be constructed for each form, explicitly covering the relevant contexts. Once the homogeneity of such scales has been established, the experimenter is free to utilize one or more of these scales in basic research on that variable. When several scales are used, separate analysis for each one would seem desirable, in addition to analyses of any composite scores obtained by conceptually appropriate combinations of certain scale scores. Empirical tests of this strategy are being initiated.

The objection may also be made that a structured test with homogeneous content will often enable the subject to see readily what variable is being measured. This possibility is especially marked for personality questionnaires. Any undesirable effects from such obviousness can be somewhat reduced by the usual practice of administering the items from several scales in mixed order within a single instrument.

The more obvious the content, the greater the potential contribution of such dispositions as the tendency to be defensive or to present oneself in a socially desirable light. The problem of assessing and controlling such tendencies is beyond the scope of this paper. To the extent that they are uncorrelated with the variable for which the test is designed, they will reduce homogeneity and the indices discussed above. If these dispositions are substantially correlated with the target variable or if they contribute most of the systematic variance, the test may appear fairly adequate in terms of the proposed criteria. Such undesirable conditions should, however, be detectable in appropriate validational studies.

Further Considerations

The Group Mean. Close examination of the data has failed to reveal any systematic contribution of the grand mean for all cells in the data matrix. To be sure, for most of the scales that have been analyzed, the mean falls near the middle of the possible range of scores. There is a slight suggestion that scales with more extreme means tend to have relatively higher proportions of remainder variance.

Dichotomous versus Scaled Scoring of Items. As noted above, the use of dichotomous scoring introduces artifactual variance. Analysis of an ideal data matrix indicates that this contribution can be as much as one-third of the total variance. Analyses of other con-

trived homogeneous matrices not having full rectangular distributions of item and of person means yield contributions from that value down to 25 per cent.

Now completely satisfactory procedure is known for estimating this component in empirical data. One possibility is to take the item means as fixed and from them to determine the maximum possible variance of person scores (cf. Carroll, 1945). The sum of this latter quantity and the variance of the item means is then subtracted from the total variance. By this method, the estimated contribution from dichotomous scoring is typically around .06 and .07, or roughly 25 to 30 per cent of the total variance (for various dichotomously scored tests).

Another approach is to compute the minimal amount of this component for the obtained set of item and person means, i.e., the amount which would have been obtained if each person had given the response predicted by the corresponding person and item means. For a few tests with dichotomous scoring, values in the order of .06 to .11 have been obtained. These are 46 to 75 per cent of the residual variance (the higher per cent when the residual is small in absolute size) and 27 to 43 per cent of the total variance.

Neither rationale appears to be satisfactory. The former yields the value which would have been found if, contrary to fact, the subjects' responses had been such as to maximize the variance of their scores. The latter analyzes the hypothetical case in which the person and item means remain as observed even though each person's response to each item is completely consistent with the overall pattern of item means: it is obvious that if such consistency were present, if the latter condition were met or closely approximated, the variances of item and person means would be greater.

Another way to approach the matter is to ask whether tests with multi-step scoring have appreciably better indices than they would have had if they had been scored dichotomously. We have applied the standard analyses to 19 scales which had multi-step or scaled scoring and compared the indices from this original scoring with those from dichotomous scoring, dividing the scoring scale near the group mean. The median increase in the relative remainder variance (s_e^2/s_t^2) is .06. The median decrease in the proportions of variance associated with item and person means is .03 for each. While these effects are only in the order of one-tenth of the

original values, they vary considerably from one test to another, and some of the largest effects occur for scales with quite favorable indices when scored with multiple steps. Thus it appears to be desirable to use more than two scoring steps whenever the content is appropriate.

The above discussion should not be taken as suggesting that the use of three to five steps in item scoring eliminates the scoring artifact completely. To obtain a completely homogeneous data matrix with considerable item and person variance and with no artifactual variance, it would be necessary to have more scoring steps than the number of steps between the highest and lowest persons when their scores are stated in conventional terms as the sum of the item scores (rather than as the mean item score). The use of several scoring steps somewhat reduces but does not eliminate the artifactual component.

The Use of Stratified Items. Some attempt has been made to explore the possibility of improving tests by selecting from existing tests a set of stratified items, a set yielding a rectangular distribution extending over a major part of the potential range from .00 to 1.00. The results have not been encouraging. While some of the indices become more favorable in some instances, such a technique for test refinement is no panacea. The findings suggest that this procedure does little or nothing to reduce the major source of difficulty, person-item interaction. They also suggest the tentative inference that significant improvement in the several indices would not be obtained by adding further items of the same kind to a test for the purpose of extending the range of endorsement values and increasing the variance of item means.

Idiosyncrasy Scores. For each cell in the data matrix, we can determine the response expected on the basis of the row (person) mean and the column (item) mean and compare it with the observed response. Thus if the test is scored dichotomously and the expected value is above .5, the response should be 1; if below, it should be 0. If the observed response is not as predicted, we can classify it as idiosyncratic. We can then count the number of such responses for each person and call this his idiosyncrasy score. Similar scores can be obtained for the several items.

It was thought that such scores might be useful for refining tests. For example, items with high idiosyncrasy scores should be

eliminated. However, the same type of improvement can be obtained by standard methods, such as item-test correlation. In principle, such scores might be used for identifying persons who seemed to be approaching or interpreting many of the items in a highly individualistic way. Whether such identification could have any practical utility remains to be demonstrated.

There seems to be little of value in such scores. They have very little internal consistency when they are analyzed in the same fashion as the conventional response data. The sole consistent finding is a negative relationship between idiosyncrasy score and distance of the person or item mean from .5 (for tests with dichotomous scoring): especially for items, the more extreme the mean, the lower the idiosyncrasy score. This finding would appear to be one further argument for constructing tests with rectangular distributions of item means rather than with means clustering near .5.

Summary

Operations are persented for evaluating the adequacy of a test in terms of a cumulative homogeneity model. Particular emphasis is placed on the proportion of the total variance in a matrix remaining after subtraction of the variances of item means and of person means. This remainder variance includes the contribution of artifactual variance from dichotomous scoring and interactions between person and item.

Analyses of existing personality tests reveal much variation in adequacy. As guides for test construction, three hypotheses are proposed: The adequacy of a test is a direct function of (a) the degree of structuring of the task established for the subject by the instructions and the situation, (b) the degree of structuring of the typical item-stimulus, and (c) the substantive homogeneity of its items.

REFERENCES

Benton, A. L. "The Interpretation of Questionnaire Items in a Personality Schedule." *Archives of Psychology* (New York), 1935, No. 190.

Campbell, D. T. "A Typology of Tests, Projective and Otherwise." *Journal of Consulting Psychology*, XXI (1957), 207–210.

Carroll, J. B. "The Effect of Difficulty and Chance Success on Correlation between Items or between Tests." *Psychometrika*, X (1945), 1–19.

Comrey, A. L. "Factored Homogeneous Item Dimensions in Personality." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 417–431.

Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika,* XVI (1951), 297–334.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. "Theory of Generalizability: a Liberalization of Reliability Theory." *British Journal of Statistical Psychology,* XVI (1963), 137–163.

Eisenberg, P. "Individual Interpretation of Psychoneurotic Inventory Items." *Journal of General Psychology,* XXV (1941), 19–40.

Endler, N. S., Hunt, J. McV., and Rosenstein, A. J. "An S-R Inventory of Anxiousness." *Psychological Monographs,* LXXVI (1962), No. 14 (Whole No. 536).

Fiske, D. W. "Homogeneity and Variation in Measuring Personality." *American Psychologist,* XVIII (1963), 643–652.

Fiske, D. W. and Butler, J. M. "The Experimental Conditions for Measuring Individual Differences." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 249–266.

Guilford, J. P. *Psychometric Methods.* (2nd ed.) New York: McGraw-Hill, 1954.

Jourard, S. M. and Lasakow, P. "Some Factors in Self-Disclosure." *Journal of Abnormal and Social Psychology,* LVI (1958), 91–98.

Levin, J. "Three-Mode Factor Analysis." Unpublished doctor's dissertation. University of Illinois, 1963.

Loehlin, J. C. "Word Meanings and Self-Descriptions." *Journal of Abnormal and Social Psychology,* LXII (1961), 28–34.

Orne, M. T. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist,* XVII (1962), 776–783.

Schutz, W. C. *FIRO: A Three-Dimensional Theory of Interpersonal Behavior.* New York: Rinehart, 1958.

Weitzenhoffer, A. M. and Hilgard, E. R. *Stanford Hypnotic Susceptibility Scale.* Palo Alto, California: Consulting Psychologists Press, Inc., 1959.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 1, 1966

PATTERNS OF MOTIVATION FOR COLLEGE ATTENDANCE[1]

ARLENE G. COHEN AND GEORGE M. GUTHRIE
The Pennsylvania State University

THE experience of counselors and the evidence of research do not agree on the role of motivational factors in academic performance. Contact with students leaves one with the firm conviction that motives to achieve have much to do with how well students perform. Given the same level of ability one student may do well, another poorly, and the difference seems clearly related to emotional or motivational differences in the students. Because this observation is made every day there have been many attempts to include measures of motivation, personality, and adjustment in prediction batteries. However, the improvement in prediction with the addition of non-intellective predictors has rarely exceeded .05, although there was room for improvement since intellective predictors produce an average multiple correlation of +55. (Fishman, 1962)

It is our impression that the addition of non-intellective variables to prediction formulas has not improved prediction appreciably because the relationships between variables assumed by multiple regression equations are not the same as those assumed by clinicians. Multiple regression equations presuppose linear compensation whereas the clinician operates with a sort of typology in which he sees individuals in certain categories rather than at certain points along continuum. As Meehl's summary (1954) suggests, however, clinicians have not demonstrated any remarkable ability to capitalize on their more flexible model.

[1] This report is based on a thesis supervised by the second author and submitted by the first author to the Pennsylvania State University, 1964.

In spite of the failure of multiple regression procedures to make productive use of motivational factors there are alternatives to the crystal ball. Stern, Stein and Bloom (1956) have outlined and evaluated a number of methods of coping with the multidimensionality of the criterion and have suggested improved methods of combining predictor variables. Saunders (1956) has developed a mathematical expression for moderator variables, variables which, although uncorrelated with either predictor or criterion, enhance the level of prediction by separating less readily from more readily predicted subjects. Middleton and Guthrie (1959) have domonstrated that there are different personality syndromes among students who have achieved at the same academic level. Taking these studies together suggests that a desirable tactic would involve categorizing subjects and developing separate prediction equations for each homogeneous group. This would result in the use of different beta weights and it would lead to the use of different predictor variables as well.

In this research we are concerned with motivations for attending college which we feel are fundamental among non-intellective determiners of achievement. There have been many studies of motives for college attendance. In one of the more extensive surveys, Iffert (1958) found an emphasis on vocational reasons, with academic, social service, personal, and family tradition considerations following in approximately that order. In their review of motivational factors, Douvan and Kaye (1962) have emphasized vocational preparation especially for males, and social mobility. They point out that motivations differ with different social levels. Havighurst (1960) has also emphasized the role of social status striving as a factor in college attendance. However, there has been little systematic study whether different patterns of motivation lead to different levels of achievement given the same level of ability. The motives for attendance and the drives for achievement have been examined in detail. But there has been little ingenuity in combining these with intellective factors to obtain improved prediction.

Since high school grades reflect the unique combination for each individual of intellective and non-intellective factors, it is not surprising that they remain the best predictors of college performance. Bloom and Peters (1961) controlling the error variance due to differences between high schools, and between colleges as well,

have obtained cross-validated correlations above .70 between high school and college performance. Reducing the variability in social influences by sampling permits an improvement of prediction. It may also be possible to improve prediction by accounting for social and motivational factors by measurement and classification.

Method

The purpose of this research was to identify patterns of motivation among male undergraduate students. It was hypothesized that there is more than one pattern of motives and that groups of students could be identified with more or less homogeneous stated sets of purposes. We sought to identify these groups by inverse factor analysis, a method in which persons are correlated with one another and the factors derived are hypothetical modal persons, each one representing a sort of type characterized by a more or less unique configuration of motives.

The correlations were based on the *Ss'* response to a 105 item *Q*-sort with a forced normal nine-point distribution. The 105 items were selected from Iffert (1957) and from the statements of students made in response to a question about their reasons for going to college. A preliminary list of 140 items was administered in a 6-point form to 64 male undergraduates. The list was reduced to 105, shown in Table A,[2] by deleting items which were frequently omitted or which showed little variability between *Ss*. Each item contained a positive statement of one purpose for attending college.

The final form of 105 items was typed on individual cards and each *S* was asked to sort so as to conform to a 9-point normal distribution. The correlation between each pair of *Ss* was computed over the 105 items.

Our *Ss* were 200 male undergraduate students enrolled in Introductory Psychology representing many fields of major interest with a concentration in business and engineering. One-half of the *Ss* were beyond their freshman year. The *Ss* were divided into two groups of 105 and 95, with the second group used to validate the findings for the first group.

A 105 $\times$ 105 matrix of correlations was generated. Ten factors, the number was our best estimate of the number of interpretable factors we might expect, were extracted and rotated orthogonally

using the Varimax solution. Correlations were computed between each item and each factor and the factors were interpreted by selecting those items with the highest correlation with each factor. Parallel data from the second 95 × 95 matrix were used to evaluate the stability of the findings on the first group of *S*s.

The naming and interpretation of factors is a matter of judgment, particularly where inverse factor analytic procedures are used. This is especially the case when the variables are people who steadfastly refuse to appear in factorially simple form. The degree to which the replication produces similar factors is also a matter of judgment.

Results

The first factor accounted for 29 per cent of the variance, the second for 9 per cent, and all ten accounted for 59 per cent of the total variance. On the replication the pattern was essentially the same with the factors accounting for 58 per cent of the total variance. The two 10 × 105 matrices of correlations between items and factors are generated. Items correlating above ±.30 with each factor are shown in Tables B-U.[2]

The results indicated that many *S*s had significant loadings on more than one factor. The number of *S*s loading .30 or more on a factor varied from 5 to 45. Communalities varied from .40 to .80 with the majority falling between .55 and .75. The rotated factors were essentially unipolar with the exception of an occasional *S* who showed a negative loading on a factor when all other *S*s who had significant loadings were positive. When items are correlated with factor loadings we find both negative and positive correlations suggesting that a factor is characterized by the acceptance of certain values associated with the rejection of other values. This will become clear in the following example.

In order to clarify our interpretation procedures we have shown in Tables 1 and 2 the items which correlate highly with Factors I and II.

Reading the items which correlate with Factor I gives one the pic-

[2] Tables A-U have been deposited with the American Documentation institute. Order Document No. 8571, remitting $2.00 for 35 mm microfilm or $3.75 for 6 by 8 in. photocopies. Order from ADI Auxiliary Publication Project, Photoduplication Service, Library of Congress, Wash. 25, D. C. Make checks payable to Chief, Photoduplication Service, Library of Congress.

TABLE 1

Items Correlating with Factor I Desire for Intellectual Development with Denial of Interest in Financial and Social Gain

Item	Correlation	Content
		Positive Correlations
34	.59	I want to broaden my overall viewpoint.
21	.58	I want to learn to think better.
5	.53	I want to find out about certain areas of knowledge.
101	.49	I want to be of service to society.
37	.48	I want to get the answers to many questions.
102	.48	I want to stimulate my interest in learning.
6	.47	I want to learn to think creatively.
45	.44	I want to know more about the world I live in.
17	.42	I want to go so that I can fulfill my purpose in life.
8	.42	I want to develop my abilities further.
32	.40	I want an education so that I can be more of a credit to my community, state and nation.
38	.45	I want to learn to understand other people better than I do now.
		Negative Correlations
99	−.63	I want to earn more money than the average person.
36	−.60	I want a college education so that I can earn a living more easily.
95	−.59	I want to go to college because attendance at college makes it easier to get a job.
72	−.59	I want a college education so that I can lead an easy life.
75	−.55	I want to avoid the hard life uneducated people live.
55	−.54	I want the respect that people give to a college educated person.
58	−.54	I want to appear educated to others.
2	−.49	I think that going to college is the "thing to do."
4	−.47	I want to go to college because my parents have always wanted me to go.
105	−.47	I want to be able to get a job that will make me important.
57	−.45	I want to prepare myself for a better paying job than I could get without going to college.
103	−.44	I want to reach a higher place in society than my parents have.

ture of an individual who gives high priority to intellectual growth and at the same time insists that considerations of financial and social gain are of low importance to him. Factor II identifies a group of individuals who are greatly influenced by parental and social expectations and who give low priority to intellectual matters which were most important to *S*s on Factor I. This procedure was followed in the identification and interpretation of subsequent factors.

A brief description of each factor, which includes the items

TABLE 2

Items Correlating with Factor II, Desire to Satisfy Parental and Social Demands and To Have Future Security with Denial of Intellectual Motivation

Item	Correlation	Content
		Positive Correlations
44	.57	I want to go to college because my parents have always wanted me to go.
104	.50	I want to go to college because my parents think I'd be wasting a good opportunity if I didn't.
88	.47	I want to go to college because the people who have influence on me think it is the proper thing to do.
95	.45	I want to go to college because attendance at college makes it easier to get a job.
36	.42	I want a college education so that I can earn a living more easily.
2	.42	I think that going to college is the "thing to do."
22	.42	I want a college education because it is very important today.
35	.40	I want a college education because it will be a requirement for most jobs in the future.
41	.38	I want to be able to provide financial aid for my parents in their old age.
91	.36	I want to go to college so that I will have a better chance for future happiness.
		Negative Correlations
37	−.55	I want to get the answers to many questions.
83	−.50	I want to specialize in one particular field in which I am very interested.
1	−.49	I want to learn for the sake of learning.
43	−.45	I want to further my education.
39	−.44	I want to meet people who have intellectual and cultural interests.
15	−.43	I want to satisfy my serious intellectual curiosity.
100	−.42	I want to associate with nationally known scholars.
76	−.40	I want to associate with college professors.
5	−.39	I want to find out about certain kinds of knowledge.
17	−.36	I want to go so that I can fulfill my purpose in life.

with highest positive and negative correlations, is given in Table 3. Both the analysis and replication are presented with an indication of the factors which we feel are confirmed. Although the rotation was an orthogonal solution, a number of factors appear to be similar. In most instances however the similarity is reduced by considering the motives which those high on the factor deny.

The matching of factors from the replication is similarly a matter of judgment. Factors from the two analyses are paired when they have many items in common or when the implications of the items appear to be identical. In this analysis, Factors I, II, V, VII,

TABLE 3

Factors from Original Analysis and from Replication with Items Showing Highest Positive and Negative Correlations

Factor	Original Analysis Name	Factor	Replication Name
I	Desire for intellectual development with denial of interest in financial and social gain. 34. I want to broaden my overall viewpoint. (+.59) 99. I want to earn more money than the average person. (−.63)	I-R	Same as factor I 8. I want to develop my abilities further. (+.61) 57. I want to prepare myself for a better paying job than I could get without going to college. (−.43)
II	Desire to satisfy parental and social demands and to have future security with denial of intellectual motivation. 44. I want to go to college because my parents have always wanted me to go. (+.57) 37. I want to get the answers to many questions. (−.55)	II-R	Same as Factor V 75. I want to avoid the hard life uneducated people live. (+.66) 47. I want to gain an understanding of the intellectual and cultural heritage of mankind. (−.50)
III	Desire for vocational preparation and service with denial of interest in personal development and pleasure. 82. I want a college degree because it is a necessity for the kind of work I want to do. (+.53) 30. I want to share in the fun that is part of college life. (−.54)	III-R	Same as Factor I 90. I want to learn more about living. (+.55) 54. I want to go to college so that I will have security in the future. (−.60)
IV	Desire for academic development with denial of interest in increasing social status. 16. I want to go to college because it is a challenge. (+.46) 42. I want to be able to be a leader in civic affairs. (−.45)	IV-R	Desire for personal development with denial of intellectual interests and need to achieve. 53. I want to learn how to get along better with other people. (+.58) 9. I want to go to the top in my field. (−.50)
V	Desire for economic gains with denial of intellectual interests. 54. I want to go to college so that I will have security in the future. (+.49) 11. I want to learn about important people. (−.50)	V-R	Same as Factor II 4. I want to go to college because my parents have always wanted me to go. (+.54) 48. I want to marry a college graduate. (−.45)

TABLE 3—Continued

Factor	Original Analysis Name	Factor	Replication Name
VI	Desire for economic and social gains with denial of interest in personal development. 99. I want to earn more money than the average person. (+.45) 26. I want a college education so that I will be a better husband. (−.44)	VI-R	Same as Factor VII 17. I want to go so that I can fulfill my purpose in life. (+.48) 2. I think that going to college is the "thing to do." (−.42)
VII	Desire for leadership and service to society with denial of interest in social status. 33. I want to get a college education so that I will be a better than average citizen. (+.40) 46. I want to develop my social skills. (−.48)	VII-R	Desire to improve social life with denial of intellectual interests. 48. I want to marry a college graduate. (+.53) 1. I want to learn for the sake of learning. (−.60)
VIII	Desire for prominence through service with denial of personal interests. 65. I want to take part in extra-curricular activities. (+.38) 27. I want to associate with the kind of person I'd like to marry. (−.48)	VIII-R	Same as Factor IX 67. I want to lead my own life. (+.61) 65. I want to take part in extra-curricular activities. (−.40)
IX	Desire for independence with denial of social interests. 67. I want to lead my own life. (+.68) 86. I want to associate with the kind of person I would like to be. (−.32)	IX-R	Desire to serve society with denial of interest in vocational preparation and social life. 32. I want an education so that I can be more of a credit io my community, state and nation. (+.39) 93. I want to meet people whose goals are similar to mine. (−.41)
X	Desire for materialistic gains with denial of social and intellectual interests. 54. I want to go to college so that I will have security in the future. (+.56) 48. I want to marry a college graduate. (−.46)	X-R	Same as VIII 42. I want to be able to be a leader in civic affairs. (+.47) 77. I want to be able to afford a family. (−.44)

VIII, and IX are considered confirmed by the second study. The positive sides of these six factors have been identified by more traditional methods of interviews and surveys. In earlier studies they have been variously named. Factor I is the rather obvious intellectual motive for college attendance. Factor II is attendance resulting from a conformity to family and social pressures. Financial security is expressed in Factor V. Leadership and citizenship are prominent motives for those on Factor VII. Social prestige and prominence are important to those on Factor VIII while those on Factor IX go to college in order to achieve independence. Factor III, which emphasized vocational purposes and on which 25 *Ss* showed high loadings, was not confirmed in the second analysis.

The unusual aspects of our results are the identification of lowly valued goals in association with the sets of important goals which make up each factor. Thus, the downgrading of social and financial benefits appeared to be part of the picture where intellectual interests are paramount. Those who are characterized by conformity tend to minimize intellectual purposes. Those who emphasize vocational purposes minimize the personal pleasures of college life. The purposes which are given consistently low values for a given group are probably as important for understanding the behavior of the members of that group as are their highly valued goals. The values which members of one group deny may be equally as strongly asserted by members of another factor.

In any given academic setting the primary set of values endorsed by the faculty may or may not coincide with those of individual students. A given student may find that he gives low priority to the purposes endorsed by the institution and as a result may not perform as well as he would where his values were shared.

The results of our analyses are to an unknown degree specific to males from one university and the pool of items used. It will be necessary to have data from other samples of subjects and items before great confidence can be placed in a specific factorial solution. The problems of the number of factors and of orthogonal or oblique solutions also remain. It will also be necessary to determine whether predicting performance of these more or less homogeneous subgroups will involve changes in beta weights and/or predictors for each group.

This research has been concerned with patterns of motivation for college attendance. A 105 item *Q*-sort of statements for attending college was compiled and given to 200 male undergraduates. The sample was divided and two inverse factor analyses were done. Six of the ten factors described in the first analysis were confirmed by the second. Implications for prediction and for counseling were considered.

REFERENCES

Bloom, B. S. and Peters, F. R. *The Use of Academic Prediction Scales for Counseling and Selecting College Entrants.* Glencoe: The Free Press, 1961.

Douvan E. and Kaye, C. "Motivational Factors in College Entrance." In Sanford, N. (Ed.) *The American College.* New York: John Wiley and Co., 1962.

Fishman, J. A. "Some Social-Psychological Theory for Selecting and Guiding College Students. In Sanford, N. (Ed.) *The American College.* New York: John Wiley and Co., 1962.

Havighurst, R. J. *American Higher Education in the 1960's.* Columbus: Ohio State University Press, 1960.

Iffert, R. E. *Retention and Withdrawal of College Students.* U. S. Department of Health, Education and Welfare Bulletin, 1958, No. 1 Washington D. C., U. S. Government Printing Office, 1957.

Meehl, P. E. *Clinical vs. Statistical Prediction.* Minneapolis: University of Minnesota Press, 1954.

Middleton, G. and Guthrie, G. M. "Personality Syndromes and Academic Achievement." *Journal of Educational Psychology,* L (1959), 66–69.

Saunders, D. R. "Moderator Variables in Prediction." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XVI (1956), 209–222.

Stern, G. S., Stein, M. I., and Bloom, B. S. *Methods in Personality Assessment.* Glencoe: The Free Press, 1956.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

PERSONALITY PATTERNS IN PRE-ADOLESCENT DELINQUENT BOYS[1,2]

HERBERT C. QUAY
University of Illinois

In a series of prior studies we have attempted to determine the principal dimensions of personality associated with juvenile delinquency in samples of institutionalized adolescents. The factorial analysis of data obtained from case histories (Quay, 1964b), questionnaire responses (Peterson, Quay and Cameron, 1959; Peterson, Quay and Tiffany, 1961; Quay and Peterson, 1964), and ratings of problem behavior (Quay, 1964a) has revealed a remarkable conceptual similarity in factor structure. Generally, the majority of variance is accounted for by three factors which we have labeled psychopathic-unsocialized, neurotic-disturbed, and subcultural-socialized. A fourth factor, usually accounting for less than 10 per-cent of the variance, has also emerged on occasion and we have called this vector inadequacy-immaturity. This structure also bears considerable similarity to that suggested by other studies of problem behavior in nondelinquent as well as delinquent children. (Hewitt and Jenkins, 1946; Jenkins and Glickman, 1947; Peterson, 1961; Quay and Quay, 1965).

Since the development of a system of statistically covarying and

[1] This research was supported by grants (M-5627 and MH-08437-01) from the National Institutes of Health, U. S. Department of Health, Education and Welfare.

[2] The author wishes to express his appreciation to Mr. Harry R. Wilson, Chief, Division of Institutions, California Department of the Youth Authority and Mr. Noel G. Bonelli and his staff of the Fricot Ranch School for their generous cooperation in this study.

conceptually coherent personality dimensions relevant to delinquency has important implications for studies of both etiology and remediation the present study sought to extend the earlier work to a sample of younger delinquent boys.

Method

Subjects

Ss were 122 males institutionalized in a comparatively small state institution with a specialized program for younger delinquents. The mean age of the sample was 12.27 years with an *SD* of 1.36. *Ss* studied represented a sizeable proportion of the entire institution population and can certainly be considered representative of the total population. The institution itself handles almost all of the serious younger delinquents in the state in which it is located.

Procedure

Behavior Ratings

Ratings of behavior were obtained from two sources: classroom teachers and cottage supervisors. The rating scale, containing 58 items descriptive of problem behavior, had been used in an earlier study (Quay, 1964a) and had been originally developed by Peterson (1961). While raters were asked to differentiate "mild" from "severe" problems, in the statistical analysis ratings were simply dichotomized on a "present" or "absent" basis. A total of 221 ratings were analyzed; the duplicate ratings on 99 of the *Ss* allowed an assessment of the reliability of the factor scores obtained. Items which were noted in less than 10 percent of the cases were not further analyzed. The remaining 49 items were intercorrelated by means of Phi-coefficients, subjected to a principal axis factor analysis and rotated to the varimax criterion (Kaiser, 1958). Squared multiple correlations were used as communality estimates. On the basis of the prior research an arbitrary decision was made to rotate only three factors. Inspection indicated that among the three additional factors extracted but not rotated there were no loadings greater than .50 and only two greater than .40. In previous work we have routinely not rotated any factor not having at least one loading of .40.

Case History Analysis

The analysis of the case histories was accomplished by social workers who were familiar with the records of the *S*s. Ratings of "present" or "absent" were made on a checklist containing 36 items. Any item noted in less than 10 percent of the cases was dropped; the remaining items were analyzed in exactly the same way as were the behavior ratings. In this analysis there were only three loadings of .40 or greater among the three factors extracted but not rotated.

It should be noted that the reliability of the data contained in the case histories is of course unknown. Assessing of rater reliability of judgments from this data was also impractical within the limits of this study. However, as has been true in the previous studies of history ratings (Hewitt and Jenkins, 1946; Quay, 1964b) the results do not appear to be seriously attenuated by lack of reliability from either source.

Personality Questionnaire

*S*s were also administered a factorially derived questionnarie which had been designed to measure three personality factors from this domain previously found to be associated with delinquency (Quay and Peterson, 1964). In addition to considerable factorial purity for the subscales all of the items in the questionnaire statistically differentiate delinquents from nondelinquents. *S*s were administered this inventory in classroom groups; in the lower grades the items were read aloud to *S*s by *E*. Factor scores were obtained for each *S* and these scores were subsequently related to factor scores obtained from the other media.

Results

Behavior Ratings

The three factors rotated accounted for 65 percent of the common factor variance. Table 1 presents the rotated factor loadings in company with loadings for the same variables from two other samples. The first factor clearly has to do with overt aggression and hostility coupled with irritability, defiance and an absence of concern for others. In keeping with prior work we have labeled this

factor psychopathic-unsocialized. The second factor is obviously a dimension of anxiety, depression, inferiority and withdrawal. In accord with previous work this factor has been labeled neurotic-disturbed. Factor III is less clearly interpretable and accounts for only 9 percent of the common factor variance. The high loadings on shortness of attention, easily flustered, shyness, lack of interest, laziness and daydreaming suggest an inadequacy-immaturity reaction and the factor has been so labeled.

TABLE 1

Rotated Factor Loadings for the Ratings of Problem Behavior with Loadings from Previous Studies Presented for Comparison

	Present Study				Adolescent Delinquents			Eighth Grade Students		
Variables	I	II	III	h^2	I	II	III	I	II	III
7. Boisteriousness, rowdiness.	.69	.00	.08	.48	.71	−.14	.18	.60	−.03	.15
5. Disruptiveness, tendency to annoy and bother others	.67	.00	.15	.47	.77	−.05	.11	.70	−.12	.21
47. Irritability; hot-tempered, easily aroused to anger.	.65	.16	−.15	.47	.53	.29	.37	*		
30. Disobedience, difficulty in disciplinary control.	.64	−.03	.06	.41	.74	.11	.00	.62	.00	.22
37. Hyperactivity; "always on the go."	.60	−.05	.09	.37	.37	.03	.53	.54	.08	−.01
32. Uncooperativeness in group situations.	.59	−.04	.13	.37	.74	.15	.01	.36	−.07	.25
41. Impertinence sauciness.	.57	−.16	.02	.35	.62	.21	.29	.33	.11	.07
29. Tension, inability to relax.	.57	.28	.09	.41						
13. Jealousy over attention paid other children.	.56	.23	−.02	.37	.21	.00	.52			
20. Fighting	.55	.02	−.07	.31	.60	.06	.02			
44. Profane language, swearing, cursing.	.55	.06	−.03	.31						
40. Negativism, tendency to do the opposite of what is requested.	.53	.04	.27	.36						
21. Temper tantrums.	.52	.22	−.05	.35	.44	.25	.44			
39. Destructiveness in regard to his own and/or others property.	.51	.06	.18	.29						

TABLE 1—Continued

Variables	Present Study				Adolescent Delinquents			Eighth Grade Students		
	I	II	III	h^2	I	II	III	I	II	III
17. Inattentiveness to what others say.	.50	.22	.52	.57	.64	.24	.18	.46	−.08	.58
2. Attention-seeking, "show-off" behavior.	.49	.01	−.09	.25	.70	−.07	.20	.61	−.08	.19
38. Distractibility.	.46	.03	.53	.50	.34	.30	.62	.59	−.11	.36
1. Restlessness, inability to sit still.	.44	.03	.04	.19	.46	.14	.37	.70	−.04	.15
6. Feelings of inferiority.	.06	.65	.05	.43	.05	.47	.29	−.07	.57	.17
18. Easily flustered and confused.	.21	.60	.25	.47	.19	.55	.46	.14	.34	.02
26. Anxiety, chronic general fearfulness.	.21	.59	.07	.40	.31	.51	.19			
10. Shyness, bashfulness.	−.27	.59	.08	.43	−.28	.54	.13	−.42	.38	.00
4. Self-consciousness; easily embarrassed.	−.08	.59	.00	.35	−.05	.48	.39	.00	.54	−.11
16. Lack of self-confidence.	.03	.57	.22	.37	.12	.66	.30	−.17	.63	.22
31. Depression, chronic sadness.	.01	.55	.09	.31	.11	.61	.03			
24. Hypersensitivity; feelings easily hurt.	.21	.52	.00	.31	.24	.49	.52			
11. Social withdrawal, preference for solitary activities.	−.34	.41	.27	.36	−.06	.67	.04	−.18	.15	.12
35. Clumsiness, awkwardness, poor muscular coordination.	.01	.40	.26	.23						
46. Nervousness, jitteriness, jumpiness; easily startled.	.39	.40	.17	.34	.22	.42	.43	.28	.35	−.14
3. Doesn't know how to have fun; behaves like a little adult.	−.12	.39	.10	.16	.23	.60	.22	−.15	.31	.03
14. Prefers to play with younger children.	.17	.36	.18	.19						
8. Crying over minor annoyances and hurts.	.24	.36	.17	.22	.08	.34	.18			

TABLE 1—Continued

Variables	Present Study I	II	III	h^2	Adolescent Delinquents I	II	III	Eighth Grade Students I	II	III
25. Laziness in school and in performance of other tasks.	.24	.09	.66	.49	.55	.20	.00	.22	−.15	.59
28. Masturbation.	−.01	.27	.62	.46	.10	.17	.63			
42. Sluggishness, lethargy	−.16	.29	.58	.45	.22	.62	−.05	−.17	.02	.27
43. Drowsiness.	−.07	.25	.52	.34						
15. Short attention span.	.37	.19	.55	.47	.59	.11	.28	.44	.11	.56
9. Preoccupation; "in a world of his own."	−.21	.18	.48	.31	.14	.60	.28	−.04	.21	.62
19. Lack of interest in environment, generally "bored" attitude.	−.03	.29	.47	.31	.49	.49	−.02	−.02	−.02	.48
27. Excessive daydreaming.	.41	−.07	.50	.42	.08	.70	.26	.13	.20	.57
12. Dislike for school.	.31	.17	.40	.28						
22. Reticence, secretiveness.	−.04	.14	.25	.08				−.23	.18	.08
23. Truancy from school.	.04	−.03	.05	.00						
33. Aloofness, social reserve.	−.30	.20	.21	.17	−.01	.29	.02	−.32	.22	.20
34. Passivity, suggestibility; easily led by others.	.12	.28	.27	.17	.47	.21	.30	.13	.14	.38
36. Stuttering.	.18	.12	.16	.07						
45. Prefers to play with older children.	.03	.00	−.03	.00						
48. Enuresis, bedwetting.	.10	.08	.08	.02						
49. Specific fears, e.g., of dogs, of the dark.	.08	.19	.09	.05						
% common variance	35	21	9		49	17	7	51	23	12

* Where no loadings appear for the earlier studies the particular variable was not analyzed due to infrequency in the sample.

By inspection the comparability between the results of this study and those of previous studies on the samples of both delinquents (Quay, 1964a) and public school students (Quay and Quay, 1965) is quite good. This improvision is well substantiated, at least for the

first two factors, by the pattern of Tucker coefficients of factor similarity[3] which may be found in Table 2.

TABLE 2

Coefficients of Factor Similarity for Loadings in Table 1. "Validity" Values Underlined.

	Adolescent Delinquents			Eighth Grade Students		
Factors	I	II	III	I	II	III
Present Study						
I. Unsocialized-Psychopathic	<u>.87</u>	.27	.48	<u>.93</u>	−.06	.45
II. Disturbed-Neurotic	.23	<u>.86</u>	.63	−.09	<u>.72</u>	.27
III. Inadequate-Immature	.62	.62	<u>.49</u>	.26	.29	<u>.89</u>

Case History Analysis

Four factors accounted for 66 percent of the common factor variance. Rotated factor loadings may be found in Table 3 in company with loadings for the same variables obtained in the earlier study of institutionalized adolescents (Quay, 1964b). Factor I seems a clear representation of a dimension of aggression and overt,

TABLE 3

*Rotated Factor Matrix for Case History Variables**

	Present Study Factor					Adolescent Delinquents Factor				
Variable	I	II	III	IV	h^2	I	II	III	IV	h^2
Assaultive	59	−05	−05	−03	35	54	14	−16	−03	34
Has bad companions	00	73	08	−09	55	06	59	−02	−08	36
Seclusive, stays to himself	−34	−28	38	−03	34	−14	−38	07	08	17
Initiates fights	68	−03	−11	−07	48	−04	58	−16	−13	33
Shy	−45	−11	58	10	56	−11	−33	46	04	33
Cruel	43	−11	−02	−12	21					
Engages in co-opertive stealing	−10	63	03	−36	54	−06	15	−02	41	19
Apathetic, emotionless	−09	−02	42	07	19					
Quarrelsome	59	09	03	09	37	43	−09	−12	−20	25
Loses interest quickly	10	−08	27	−17	12	−05	−05	−08	13	03
Defies authority	46	12	−04	32	33	47	14	−26	06	31
Engages in furtive stealing	07	−08	13	43	21	18	−25	−15	02	12

[3] Ledyard R Tucker, personal communication. This coefficient is obtained by dividing the sum of the cross-products of the loadings by the square root of the product of the two sums of squares. It is essentially a correlation coefficient, uncorrected for origin.

TABLE 3—Continued

Variable	Present Study Factor I	II	III	IV	h^2	Adolescent Delinquents Factor I	II	III	IV	h^2
Worries	01	12	47	−03	24	−07	−13	67	04	47
Engages in malicious mischief	34	12	37	−04	27	01	00	−21	06	05
Habitually truant from school	−17	−07	−23	05	09	17	23	04	11	09
Sensitive	−05	00	34	−04	12	13	−04	70	−01	51
Unable to cope with a complex world	14	−19	49	47	52	04	02	06	55	30
Timid	−43	−05	56	06	50	−09	−17	59	15	41
Has inadequate guilt feelings	25	14	−04	44	28	14	−10	−27	−14	12
Habitually truant from home	−06	−13	−25	−01	08	−27	27	−21	42	37
Submissive	−32	−01	34	03	22					
Stays out late at nights	−01	52	−13	05	29	−05	48	−26	25	36
Irritable	48	00	20	−15	29	61	03	10	00	32
Accepted by a delinquent subgroup	09	73	−01	08	55	−15	63	09	−32	53
Lonesome	−13	−32	34	21	28					
Verbally aggressive, impudient	61	22	−09	−09	44	56	−04	−19	00	35
Strong allegiance to selected peers	08	61	22	−05	43	13	36	−07	−29	24
Incompetent, immature	05	−18	29	25	18	08	−20	02	36	17
Obscene, uses foul language	49	14	01	12	27					
Feels persecuted, believes others unfair	29	00	11	−02	10	49	16	00	15	29
Has anxiety over own behavior	−24	03	11	−02	07	−07	−23	42	01	23
Callous, little concern for others	22	18	−02	−38	23	16	16	−26	−06	12
Unable to profit by either praise or punishment	20	−12	14	−37	21	31	−28	−31	−03	27
Suspicious, trusts no one	24	−10	−07	15	10	00	−12	−19	−12	06
Has engaged in sex delinquencies	−09	16	−05	26	10					
% common variance	27	17	14	8		26	15	10	9	

* decimals omitted

unbridled hostility and in keeping with prior studies has been labeled psychopathic-unsocialized. The second factor reflects a syndrome of gang activities and as in the past has been labeled unsocialized-subcultural. Factor III contains elements of withdrawal, submissiveness and anxiety and appears to warrant the application

of the disturbed-neurotic label. The fourth factor is again less well defined but its components suggest the inadequacy-immaturity label applied in the earlier studies. Coefficients of factor similarity have been calculated and these may be found in Table 4. These coefficients indicate a high degree of similarity in the structure obtained from the two samples for all but factor IV.

TABLE 4

Coefficients of Factor Similarity for Loadings in Table 3. "Validity" Values Underlined.

Present Study	Adolescent Delinquents Factor I	II	III	IV
I. Psychopathic-Unsocialized	.73	.30	−.46	−.17
II. Socialized-Subcultural	.08	.65	−.10	−.20
III. Neurotic-Disturbed	.02	−.44	.57	.38
IV. Inadequacy-Immaturity	.01	−.13	−.01	.14

Factor Scores

Table 5 presents the means and standard deviations for the factor scores from the three domains. These values are comparable to those obtained in the earlier researches on institutionalized adolescents.

TABLE 5

Means and Standard Deviations for Factor Scores.

Domain and Factor	Mean	SD
Questionnaire		
Psychopathic-Unsocialized	12.22	8.10
Neurotic-Disturbed	15.80	5.52
Socialized-Subcultural	18.01	3.27
Behavior Rating (Teacher)		
Psychopathic-Unsocialized	7.09	5.18
Neurotic-Disturbed	5.59	4.05
Inadequate-Immature	4.42	3.27
Behavior Rating (Cottage)		
Psychopathic-Unsocialized	8.15	5.07
Neurotic-Disturbed	6.31	3.31
Inadequate-Immature	4.15	2.60
Case History		
Psychopathic-Unsocialized	1.96	2.49
Neurotic-Disturbed	1.50	1.48
Inadequate-Immature	1.23	1.22
Socialized-Subcultural	1.43	1.62

As noted above 99 of the *S*s had been rated by both teachers and cottage supervisors. Correlations between these two sets of ratings were quite low; .38 for Factor I, .20 for Factor II and .30 for Factor III. These reliabilities are even poorer than those found previously (Quay, 1964a); nevertheless while there is obviously both situational and rater variance (in addition to considerable restriction of range) the raters are using a common dimensional system.

Factor Equivalence Across Domains

In order for a factor analytically derived descriptive system to be most meaningful some degree of equivalence of conceptually similar factors evolving from different domains ought to be empirically established. The establishment of such linkage has been a point of contention and a serious problem for factor analytic research on personality (Becker, 1960; 1961; Cattell, 1961a; 1961b).

We have approached the establishment of empirical validity for cross-domain comparisons by obtaining factor scores for each *S* on all factors, intercorrelating these scores, and presenting the resulting intercorrelations (Table 6) in a multitrait-multimethod matrix (Campbell and Fiske, 1959).

On the whole the validities are poor and the within method correlations are frequently higher for different traits than the between

TABLE 6

Correlations among Factor Scores for Questionnaire, Behavior Ratings and Case History Data "Validity" Values Underlined.

		Questionnaire			Domain Behavior Ratings*			History			
Factors		N	P	S	N	P	I	N	P	I	S
Questionnaire	N										
	P	.30									
	S	.49	.02								
Ratings	N	.11	.22	−.03							
	P	.08	.16	.03	.22						
	I	.04	.31	−.06	.57	.37					
History	N	−.07	−.01	.00	.19	−.26	.07				
	P	.15	−.08	.18	−.03	.33	−.09	−.41			
	I	−.03	−.03	.03	.01	−.14	−.01	.35	.00		
	S	−.17	−.04	.01	−.17	−.04	−.20	.05	.05	−.09	

* Based on the averages of values obtained from teachers and cottage supervisors.

correlations for the same traits. It is clear that the test factors have little empirical validity for the differential prediction of any of the factors from the other domains. However, for both psychopathy and neuroticism there is some suggestion of cross-domain matching between the history analysis and the behavior ratings; here the correlations between cross-domain factors are higher for the same trait than for the others.

Discussion

A number of conclusions seem appropriate. When one considers data from within either the case history record or behavior rating domains the structure similarity across varying samples is striking. This is true not only for other samples of delinquents but for public school children as well. Thus, it appears that the primary dimensions of problem behavior as manifest within each domain are common to all children.

When cross-domain relationships are considered convergence is much less clear. One can speculate about the causes of this lack of convergence and this has been done at length (Cattell, 1961a). In this instance it is clear that the institution setting (from which come the behavioral ratings) and the prior life setting (from which come the history data) are situational contexts separated temporally as well as spatially. Further there are reductions in the magnitude of the relationships resulting from the known low reliability of the behavior ratings; poor reliability likely characterizes the case history data also.

The self reports about behavior and attitudes required by the test are certainly not highly related to the judgments about conceptually similar behavior and attitudinal dimensions provided by others. Here again both situational and psychometric factors are at work. In addition the problem of defensiveness and other response biases on the part of the *S*s must be considered.

In view of the ample opportunities of operation for both situational and more purely psychometric influences in the cross-domain comparisons we are inclined to feel that the present results may be all that can be expected and even the limited magnitude of the cross-domain relationships argue for the presence of basic trait dimensions which are observable in all three domains. On the basis of this research and the earlier studies these dimensions now appear

to be well enough established to suggest that their origin and maintainence are of considerable importance in the understanding of juvenile delinquency.

REFERENCES

Becker, W. C. "The Matching of Behavior Rating and Questionnaire Personality Factors." *Psychological Bulletin,* LVII (1960), 201–212.

Becker, W. C. "Comments on Cattell's Paper on 'Perturbations' in Personality Structure Research." *Psychological Bulletin,* LVIII (1961), 175.

Campbell, D. T. and Fiske, D. W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin,* LVI (1959), 81–105.

Cattell, R. B. "Theory of Situational, Instrument, Second Order, and Refraction Factors in Personality Structure Research." *Psychological Bulletin,* LVIII (1961), 160–174. (a)

Cattell, R. B. "Cattell Replies to Becker's 'Comments.'" *Psychological Bulletin,* LVIII (1961), 176. (b)

Hewitt, L. E. and Jenkins, R. L. *Fundamental Patterns of Maladjustment, the Dynamics of Their Origin.* Springfield: State of Illinois, 1946.

Jenkins, R. L. and Glickman, Sylvia. "Patterns of Personality Organization among Delinquents." *Nervous Child,* VI (1947), 329-339.

Kaiser, H. F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika,* XXIII (1958), 187–200.

Peterson, D. R. "Behavior Problems of Middle Childhood." *Journal of Consulting Psychology,* XXV (1961), 205–209.

Peterson, D. R., Quay, H. C., and Cameron, G. R. "Personality and Background Factors in Juvenile Delinquency as Inferred from Questionnaire Responses." *Journal of Consulting Psychology,* XXIII (1959), 395–399.

Peterson, D. R., Quay, H. C. and Tiffany, T. C. "Personality Factors Related to Juvenile Delinquency." *Child Development,* XXXII (1961), 355–372.

Quay, H. C. "Personality Dimensions in Delinquent Males as Inferred from the Factor Analysis of Behavior Ratings." *Journal of Research in Crime and Delinquency,* I (1964), 33–37. (a)

Quay, H. C. "Dimensions of Personality in Delinquent Boys as Inferred from the Factor Analysis of Case History Data." *Child Development,* XXXV (1964), 479–484. (b)

Quay, H. C. and Peterson, D. R. "The Questionnaire Measurement of Personality Dimensions Associated with Juvenile Delinquency." Unpublished paper, 1964.

Quay, H. C. and Quay, Lorene C. "Behavior Problems in Early Adolescence." *Child Development,* XXXVI (1965), 215–220.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

SOCIAL DESIRABILITIES AMONG MEASURES OF SOCIAL DESIRABILITY

BERNARD SPILKA, JOHN HORN and LEONARD LANGENDERFER
University of Denver

About two decades ago Cattell (1944) proposed a broad logical distinction between the behavior persons exhibit under instructions to answer questions about themselves and that revealed when requested to find the "best" answer or otherwise describe or classify stimuli. He suggested that responses gathered in the first-mentioned manner be termed *Introspective, self-estimate,* or *Q-data,* and that behavior observed in the latter way be called *Objective-test* or *T-data* (both to be contrasted with ratings by others or *L-data*). More recently Jackson and Messick (1958) and Torgerson (1960) have discussed essentially the same distinction under the heading of "responses" and "judgments," while Coombs (1953) drew a similar comparison in his classification of Type I and Type II models. This basic distinction is currently at the crux of much of the controversy over the place which the concept (or concepts) of social desirability should occupy in personality or general psychological theory.

Several studies (Cattell, 1957; Edwards, 1957; Heilbrun and Goodstein, 1961; Loehlin, 1961; Messick, 1960; Rozynko, 1959) have indicated that the ways in which persons answer questions about themselves are related to the ways in which they judge (perhaps value) the behavior they are instructed to attribute to themselves. Specifically, it has been shown (Edwards, 1957) that, if a respondent judges certain behavior to be "socially desirable," he is more likely to ascribe it to himself than if he considers it to be "undesirable." But it is now quite apparent from the work of

Loehlin (1961), Messick (1960), Morris and Jones (1955), Osgood, Suci and Tannenbaum (1957) and others, that there are noteworthy differences in the patterns of behavior (beliefs, etc.) people regard as desirable. Messick, for example, using a sample of 42 stimuli, found that nine factors were needed to account for the intercorrelations between judgments about the desirability of behavior described in these rather typical self-report items. Some of this work, notably that of Buss (1959), Jones and Thurstone (1955), Loehlin (1961), and Osgood (1957) suggests that a few of these dimensions of "judgment" may be closely related to differences in language usage or in the actual concepts which the language terms are used to "tag."

It has long been apparent (Cattell, 1957), however, that there are several consistent and distinct patterns of behavior which people are likely to accept or reject in self-ascription. Edwards (1957) has documented a case for the contention that a substantial proportion of the covariation among such self-reports can be accounted for by a general set of cultural standards in relation to which most personal and social behavior can be assessed as desirable or undesirable. Granting that this general dimension represents a potentially useful concept, it must be noted that it does not account for all of the reliable variance of lower order factors (Horn, 1963).

In sum, whether these phenomena are approached through the methodology of objective tasks, or via subjective self-assessments, the evidence suggests that (1) there are probably *several* noteworthy dimensions of individual differences in conceptions of what are and what are not socially desirable characteristics of people, (2) there are probably also several significant dimensions of individual differences in self-evaluations, (3) lawful relationships between the dimensions of (1) and (2) probably exist, (4) there is a definite need to establish more firmly the salient dimensions in both (1) and (2), and to specify precisely the correspondence between these dimensions. The present study represents one attempt to deal with these questions, particularly with (4).

Method

Seven putative measures of a social desirability attribute of individual differences were studied to determine first, if the hypothesis that all can be regarded as operationally representing a single

concept could be retained and, second, if this hypothesis had to be rejected, the number and nature of dimensions reliably assessed by these measures.

Subjects

A complete set of responses for all measures was obtained from each of 50 high school seniors.

Measures

The scales here studied, all purported to be measures of a social desirability "set," are classified in Table 1 as either *Q*-data or *T*-data devices, following as nearly as possible the criteria set down by Cattell (1958) and Scheier (1958). The scales were presented to all *S*s in the orders specified in a 7 by 7 Latin Square. The rationale for use of each scale is discussed below.

TABLE 1

SD Measures

T-Data	*Q*-Data
Desoto "Social Approval" (DSA)	Bills Perceived Self (BPS)
Borislaw "Perfect Person" (BPP)	Desoto Well Being (DWB)
Bills Ideal Self (BIS)	Edwards Social Desirability (ESD)
	Marlowe-Crowne Social Desirability (MCSD)

The Edwards "scale" contains items selected because the behavior implied by the keyed response was consistently rated as highly desirable by the judges employed in the test construction studies (Edwards, 1957). As here used, the scale requires the subject to endorse the item statement for the behavior in question relative to himself.

Crowne and Marlowe (1960) argued that the content of the Edwards scale, based on items drawn from the MMPI, is overly resrticted to questions about symptoms of mental illness and is, therefore, primarily a measure of a desire to avoid labeling oneself as "sick." Using methods similar to those employed by Edwards, they developed a self-rating instrument, the Marlowe-Crowne *SD* scale (MCSD), containing few items which involve pathological content.

The Bills Index of Adjustment and Values (IAV) (Bills, 1958)

was originally designed to provide three measures, the two employed here and the Perceived Self Scale which instructs the subject to rate himself as he in fact is, while the other, called the Ideal Self Scale (BIS), asks the respondent to make his ratings in accordance with the kind of person he would ideally like to be. A judgment scale for desirability of these items was found (Cowen and Tongas, 1959; Spilka, 1961) to be nearly a perfect linear transformation of that based on the Edwards Social Desirability Scaling Item Procedure. The original scales were, therefore, interpreted as largely measuring social desirability.[1]

The Borislaw (1958) "Perfect-Person" measure (BPP) derives from the theory and procedure used by Ruch (1942) that a social desirability "set" will be indicated by the responses of subjects instructed to answer questions in the way they believe the "perfect person" would answer. Since this requires the *Ss* to rate an external "object," the measure is seemingly objective by Cattell's definition.

A similar rationale is given for the "Social-Approval" measure (DSA) developed by DeSoto, Kuethe, and Bosley (1959). Here the subject is instructed to respond in a way that would, if the response were examined by others, gain him social approval. These authors also provided a scale of "well being" (DWB) thought to be confounded by social desirability and here classified as a *Q*-data measure.

Procedures and Analysis

One of the best developed linear models for examining the first hypothesis stated above is still that created by Spearman (1904) and his students (Spearman, 1927). This embodies a stringent set of criteria, but allows one to retain a hypothesis with considerable confidence if the requirements of the model are satisfied. In essence it demands that scales measure *one and only one* attribute in common. The total variance of the measures need not be accounted for by this one factor, since the model explicitly allows that measures

---

[1] It should be noted, however, that when the subjects in Spilka's (1961) study were instructed to endorse the items to describe themselves or their ideal selves, the measures correlated .36 and —.09 respectively with the self-ratings of ESD. These results thus rather clearly support the position taken by Thurstone and Chave (1929) over 30 years ago to the effect that there is no necessary relationship between scales derived by judgmental approaches and scales of endorsement.

may be unreliable or may reliably measure attributes not also measured by other scales in a battery, but all common variance, as represented by the intercorrelations, will be exhausted by a single dimension.

Examination of the intercorrelations among our measures of social desirability (Table 2) reveals that the requirements of the Spearman model were not satisfied by these data, although positive manifold, indicating the presence of one (but not only one) general factor in a hierarchical solution (Schmid and Leiman, 1957) is definitely suggested.

TABLE 2

SD Intercorrelation Matrix (Below Diagonal) Residuals After Extraction of 3 Factors (In and Above Diagonal), and Spearman "g" Loadings

	DSA	BPP	BIS	BPS	DWB	ESD	MCSD
DSA	−001*	−000	001	−001	−001	010	−004
BPP	684	−005	026	−002	004	−009	017
BIS	686	584	002	000	−003	026	−029
BPS	684	567	652	−003	024	−029	024
DWB	516	596	336	407	000	−002	−007
ESD	118	144	040	141	143	031	−018
MCSD	033	095	−072	146	067	412	007
"g" loading	455	519	252	650	331	226	022

* Correlations and residuals are here recorded to only three places (with decimal points omitted), although actual calculations carried ten places throughout.

Since the hypothesis of one factor could not be retained, a multiple-factor hypothesis was explored by use of the methods of principal axes factor extractions with reduced communality estimates (Harman, 1960). Unities were inserted in the principal diagonal of the matrix of correlations and all principal factors were extracted. The latent root of the fourth factor proved to be less than 1.0, while that of the third factor was 1.03. Hence, following the proofs and arguments of Guttman, Kaiser, and Tryon (Harman, 1960), it was estimated that three common factors could be reliably determined. Five centroid iterations were carried out to estimate the reduced communalities (entered in the fourth column

of Table 3), these were inserted in the diagonals of the correlation matrix and three principal axes factors were computed. The residuals, given in and above the principal diagonal in Table 2, were all essentially zero at this point, the largest being 0.031. The results from this analysis are given in Table 3.

TABLE 3

Principle Axes Factors and "Entered With" h^2

	I	II	III	h^2
DSA	859	−101	040	749*
BPP	806	000	−199	683
BIS	744	−225	213	623
BPS	799	009	279	738
DWB	682	055	−431	583
ESD	197	620	029	455
MCSD	116	654	083	455
ROOT	3.03	1.88	0.36	
% COMMON VARIANCE	70.1	20.4	8.4	

* These are the reduced communalities which resulted from five centroid iterations. They will differ slightly from the sums of squares across factors.

As would be expected from an inspection of the intercorrelations, these results show that a major portion of the common variance (and, indeed, the major portion of the reliable variance of the first five variables) is accounted for by the first factor and that the ESD and MCSD form a substantial doublet factor largely independent of the first. The third factor would probably not be identified by inspection of the correlations alone. It is of borderline significance by Kaiser's (1958) test, as noted above and hence will not be emphasized in our later discussions.

The first factor involves all of the objective measures, suggesting high evaluation of self and a claimed sense of well-being. The self-evaluations made in the Edwards and Marlowe-Crowne devices, form a quite distinct factor, although low objective ratings of the "ideal self" also produce some variance.

In an attempt to retain the hierarchical order of the dimensions, while improving their interpretability and increasing the likelihood of invariance, the factors were rotated to that approximation to simple structure which is given by Kaiser's (1958) Varimax Criterion. These results are presented in Table 4.

The first two factors are not much altered by this rotation, al-

TABLE 4

Principal Axes Factors Following Varimax Rotation

	I	II	III
DSA	753	043	426
BPP	564	105	600
BIS	773	−007	217
BPS	807	169	194
DWB	284	104	702
ESD	056	642	090
MCSD	011	670	002

though the well-being measure loads substantially less in the first and "ideal self" ratings sink into the hyperplane of the second, but the third factor is found to be defined by a stated sense of well-being, objective ratings of the "perfect person" and the attempt to reply in a way that would earn social approval.

Discussion

A substantial proportion of the reliable variance in all measures is involved in the three factors here obtained, but the Edwards and Marlowe-Crowne devices, particularly, are not fully defined in this space. Judging from the results of previous work (Horn, 1963), these two variables would enter prominently into a general anxiety factor, although it is now also well established (Edwards, 1963; Edwards, Diers, and Walker, 1962; Horn, 1964) that, if the sampling of variables is concentrated in the Q-data medium *alone*, the ESD and MCSD will fall on somewhat distinct factors. Thus, these two variables here define what would be regarded as a second-order factor in many analyses.

From our point of view, the more interesting finding is the relationship between the objective (i.e., T-data) and subjective (i.e., Q-data) rating tasks as represented by the first factor in the principal axes or Spearman solution. This could be due to consistency in the use of scales (i.e., be an extremity or "in between" response set). More reasonable is the notion that these measures are keyed according to what, on the average, in the upper and middle classes of our society are regarded as socially desirable qualities, and that some people more nearly agree with this definition of the "good guys" than do others. This may be termed a social desirability "set"—but it must be noted that in our results this is largely independent of

the stereotype defined by the keys of the Edwards and Marlowe-Crowne devices. The latter fact may be due to the influence being observed in the recent studies of Edwards and Heathers (1962) and Edwards, Diers, and Walker (1962) in which very high correlations are demonstrated between first factor loadings on the MMPI and 61 personality scales with ESD. This suggests that contrary to intentions, the Marlowe-Crowne, like the Edwards, contains a substantial portion of variance due to pathological item content. With loadings as low as those here observed in Factor II, the alternative hypothesis that this factor indicates commonness between the non-pathological subcomponents of the Edwards and the Marlowe-Crowne devices cannot be rejected.

Rather than using the now obviously ambiguous term "social desirability," we would suggest at least a tentative identification of Factor I with the seemingly broader concept of self-sentiment discussed by Cattell (1957) and more fully in the reference cited. Factor II would seem better considered within the context of the concept of anxiety, a term which is no more ambiguous than social desirability and one for which the relational fertility in general psychological theory is definitely greater.

Since the third factor is of only borderline significance, we will not occupy space with a discussion of it here.

Summary

An analysis by factor analytic methods of the correlations among seven putative measures of a social desirability "response-set" revealed that (1) there is no single attribute of social desirability "set," (2) although objective measures of what are and what are not desirable qualities of persons are related to similar self-evaluation measures and define what might tentatively be interpreted as a self-sentiment factor, this dimension is largely independent of that defined by the two currently most popular devices designed to measure the "set" to respond in a socially desirable manner, viz, the Edwards and the Marlowe-Crowne scales; and (3) the factor defined by these latter measures may be tentatively identified with the second-order anxiety dimension found in previous studies.

REFERENCES

Bills, R. E. *Manual for the Index of Adjustment and Values.* Auburn: Alabama Polytechnic Institute, 1958.

Borislaw, B. "The EPPS and Familiarity." *Journal of Applied Psychology,* XLII (1958), 22–27.

Buss, A. H. "The Effect of Item Style on Social Desirability and Frequency of Endorsement." *Journal of Consulting Psychology,* XXIII (1959), 510–513.

Cattell, R. B. "Psychological Measurement: Normative, Ipsative, Interactive." *Psychological Review,* LI (1944), 292–303. (a)

Cattell, R. B. *Personality and Motivation Structure and Measurement.* Yonkers-on-Hudson, New York: World Book, 1957. (b)

Cattell, R. B. "What is 'Objective' in Objective Personality Tests." *Journal of Counseling Psychology,* V (1958), 285–289. (c)

Coombs, C. H. "The Theory and Methods of Social Measurement." In L. Festinger and E. Katz (eds.) *Research Methods in the Behavioral Sciences.* New York: Dryden, 1953.

Cowen, E. L. and Tongas, P. N. "The Social Desirability of Trait Descriptive Terms: Applications to a Self-Concept Inventory." *Journal of Consulting Psychology,* XXIII (1959), 361–365.

Crowne, D. P. and Marlowe, D. A. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology,* XXXIV (1960), 349–354.

DeSoto, C. B., Kuethe, J. L., and Bosley, J. J. "A Redefinition of Social Desirability." *Journal of Abnormal and Social Psychology,* LVIII (1959), 273–275.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden, 1957.

Edwards, A. L. "A Factor Analysis of Experimental Social Desirability and Response Set Scales." *Journal of Applied Psychology,* XLVII (1963), 308–316.

Edwards, A. L., Diers, Carol J., and Walker, J. N. "Response Sets and Factor Loadings on Sixty-One Personality Scales." *Journal of Applied Psychology,* XLVI (1962), 220–225.

Edwards, A. L. and Heathers, Louise B. "The First Factor of the MMPI: Social Desirability or Ego Strength." *Journal of Consulting Psychology,* XXVI (1962), 99–100.

Harman, H. H. *Modern Factor Analysis.* Chicago: University of Chicago Press, 1960.

Heilbrun, A. B., Jr. and Goodstein, L. D. "Consistency between Desirability Ratings and Item Endorsement as a Function of Psychopathology." *Psychological Reports,* VIII (1961), 69–70.

Horn, J. L. "Second-Order Factors in Questionnaire Data." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 117–134.

Horn, J. L. "Social Desirability, Evaluation and a Hierarchical Model for the Organization of Questionnaire Responses." Rocky Mountain Psychological Association Convention, Salt Lake City, May 1964.

Jackson, D. N. and Messick, S. J. "Content and Style in Personality Assessment." *Psychological Bulletin*, LV (1958), 243–252.

Jones, L. V. and Thurstone, L. L. "The Psychophysics of Semantics: An Experimental Investigation." *Journal of Applied Psychology*, XXXIX (1955), 31–36.

Kaiser, H. F. "The Varimax Criterion for Analytic Notation in Factor Analysis." *Psychometrika*, XXIII (1958), 187-200.

Loehlin, J. C. "Word Meanings and Self Descriptions." *Journal of Abnormal and Social Psychology*, LXII (1961), 28–34.

Messick, S. "Dimensions of Social Desirability." *Journal of Consulting Psychology*, XXIV (1960), 379–387.

Morris, C. and Jones, L. V. "Value Scales and Dimensions." *Journal of Abnormal and Social Psychology*, LI (1955), 523–535.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana, Illinois: University of Illinois Press, 1957.

Rozynko, V. V. "Social Desirability in a Sentence Completion Test." *Journal of Consulting Psychology*, XXIII (1959), 280.

Ruch, F. L., "A Technique for Detecting Attempts to Fake Performance on the Self-Report Type of Personality Tests." In Q. McNemar and Maud A. Merrill (eds.) *Studies in Personality*. New York: McGraw-Hill, 1942.

Scheier, I. H. "What is an 'Objective' Test." *Psychological Reports*, IV (1958), 147–157.

Schmid, J. and Leiman, J. M. "The Development of Hierarchical Factor Solutions." *Psychometrika*, XXII (1957), 53–61.

Spearman, C. "General Intelligence, Objectively Determined and Measured." *American Journal of Psychology*, XV (1904), 201–293.

Spearman, C. *The Abilities of Man*. New York: MacMillan, 1927.

Spilka, B. "Social Desirability: A Problem of Operational Definition." *Psychological Reports*, VIII (1961), 149–150.

Thurstone, L. L. and Chave, E. J. *The Measurement of Attitudes*. Chicago: University of Chicago Press, 1929.

Torgerson, W. S. *Theory and Methods of Scaling*. New York: Wiley, 1960.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

LIFE STATUS AND INTERPERSONAL VALUES[1]

E. K. ERIC GUNDERSON and PAUL D. NELSON
U. S. Navy Medical Neuropsychiatric Research Unit
San Diego, California

The Survey of Interpersonal Values (SIV) is a forced-choice inventory purported to measure basic motivational patterns or values. Gordon (1960, p. 3) states that SIV "is designed to measure certain critical values involving the individual's relationships to other people or their relationships to him. These values are important in the individual's personal, social, marital and occupational adjustment."

Items for the SIV were selected through factor analysis, and triads of statements were assembled so that each statement keyed on one of six scales which were labelled Support, Conformity, Recognition, Independence, Benevolence, and Leadership. Statements composing triads were equated for social desirability in order to reduce susceptibility to faking. Construction and definition of the scales are described in detail in the SIV Manual (Gordon, 1960).

Gordon (1963) has presented extensive normative data and summaries of several exploratory studies relating SIV scores to occupational effectiveness, levels of professional training, scholastic achievement, psychiatric status, and delinquent behavior as well as to cross-cultural and sex differences. While these results appear logical and informative, neither theoretical formulations nor ex-

[1] This research was supported by the Bureau of Medicine and Surgery, Department of the Navy, under Research Task MR005.12–2004, Subtask 1. Opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or as necessarily reflecting the views of the Department of the Navy or of the Naval service at large.

Appreciation is expressed to Mr. Frank A. Thompson and to Mr. Dale Trower for statistical computations used.

trapolations to relationships with other important aspects of life status have yet been offered. For example, no relationships between the SIV scales and such important life history and status variables as age, intelligence, educational level, socioeconomic or power status, social interests and activities, religious orientation, familial and cultural background have been reported.

The present study is concerned with patterns of interpersonal values, as measured by the SIV, exhibited among military and civilian volunteers for the U. S. Antarctic Research Program (Operation Deep Freeze), and, within the military segment of this population, relationships of the SIV measures to a number of personal history and status variables.

Method

Subjects

Military and civilian applicants seen in psychiatric screening for Operation Deep Freeze over a two-year period were utilized for the study. Of the Navy population 7 percent were officers and 93 percent were enlisted men. Navy applicants averaged 26 years of age and ranged from one to 23 years in naval experience; 65 percent were high school graduates. Civilian volunteers averaged 29 years of age and ranged from one to 24 years in occupational experience; 90 percent were high school graduates, and 65 percent were college graduates. Life history characteristics of Antarctic volunteers have been described in detail in previous publications (Gunderson, 1964; Nelson and Gunderson, 1963).

Procedure

Before deployment to the Antarctic each subject filled out a biographical inventory, the Personal History Booklet, and the SIV inventory. Subjects were exposed to essentially the same psychological assessment procedures both years of the study. A number of items considered to be of most importance in summarizing life history events were selected from the Personal History Booklet and related to the SIV measures.

SIV means and standard deviations were computed for the three Antarctic volunteer groups, Navy enlisted men, officers, and civilian scientists and technicians. Mean scores for these groups were

compared with each other and with data published by Gordon for a number of other groups. Intercorrelations among the SIV scales were computed for the enlisted Navy sample and compared with those published by Gordon. Finally, for the total Navy population, 13 items of life history and life status were related to the SIV scales by means of Pearson correlations and χ^2 statistics, as appropriate.

Results

Table 1 gives means and standard deviations for five normative and professional groups drawn from the SIV Manual Supplement (Gordon, 1963), and for the three Antarctic volunteer groups. The differences in mean scores shown in Table 1 suggest relationships between social status, as reflected by educational and occupational attainment, and scores on certain of the SIV scales. Conformity and Benevolence appear to be negatively related to educational-occupational status, while Independence and Leadership appear to be positively related to the same variable. Similar relationships appear to hold for the Antarctic volunteer groups on the Conformity and Independence scales. It does not seem surprising that the officers emerged as highest of the Antarctic groups on the Leadership scale; it does seem surprising that Antarctic scientists scored highest of any group on the Benevolence Scale and lowest of any group on Recognition.

The results in Table 1 suggest that Antarctic volunteers taken together have certain distinctive characteristics. Means for Conformity and Benevolence are uniformly higher for Antarctic volunteer groups than for the normative and professional groups while means for Independence and Recognition are uniformly lower for the Antarctic groups. The magnitude of differences between weighted means for the Antarctic groups versus those for the other groups reinforces the impression of a different pattern of values than is shown by other male groups.[2] Emphasis on social

---

[2] Precise tests for the significance of differences between the means of the total Antarctic population and the other groups combined were not possible since the exact variance for the combined groups was not known. However, if the largest sample variances are used as estimates of the population variances, differences between the weighted means are highly significant for all six scales. Only the differences for Conformity, Recognition, Independence, and Benevolence appear large enough to be of any practical significance, however.

TABLE 1

SIV Means and Standard Deviations for Normative and Antarctic Groups

Normative and Professional Groups:[a]	N	Support		Conformity		Recognition		Independence		Benevolence		Leadership	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
General Adults	213	15.0	5.7	14.8	6.5	11.2	5.2	16.9	7.4	15.8	5.8	16.7	7.7
H.S. Students	782	15.4	5.5	14.8	6.4	12.6	4.9	18.3	7.3	14.7	6.3	14.2	6.6
College Students	1075	14.9	5.5	12.3	6.6	12.4	5.0	19.1	7.2	13.6	6.5	17.3	7.2
Engineers	80	13.1	5.4	11.5	5.4	11.8	5.4	19.4	6.0	14.0	6.4	20.0	6.4
Physicians	10	14.6	5.9	7.3	5.3	13.3	4.1	22.6	7.2	12.2	4.6	19.9	5.4
Total (Weighted Mean)		15.0		13.4		12.3		18.6		14.2		16.2	
Antarctic Groups:[a]													
Enlisted Men	463	13.8	5.1	20.4	6.2	10.4	4.2	13.3	6.8	17.6	5.5	14.3	6.6
Officers	34	11.6	5.0	18.4	6.5	11.0	4.2	13.4	6.2	16.3	6.5	18.5	7.0
Scientists	126	13.0	5.9	15.2	7.1	9.8	4.9	16.5	7.5	18.1	5.4	17.4	6.9
Total (Weighted Mean)		13.5		19.2		10.3		14.0		17.6		15.2	

[a] Groups are ordered in terms of educational-occupational status; the relative positions of General Adults versus High School Students on this continuum are not clear from information available in the SIV Manual or Supplement.

conformity (Conformity) and helpfulness to others (Benevolence) as opposed to individual prominence (Recognition) and autonomy (Independence) would appear favorable for adaptation in small, long-term, confined groups such as those at Antarctic scientific stations.

The data of Table 1 suggested a relationship between educational-occupational status and SIV values. It seemed plausible that age, education, rank (power status), or intelligence might account in part for the observed differences among group means. Pearson correlation coefficients were computed between these presumably continuous variables and the SIV scales for the Navy volunteer population. The correlations of age with Conformity ($r = .16$) and Independence ($r = -.15$) were significant ($p < .05$) but in the opposite direction to that expected if age were to partly account for the differences in means found in Table 1. Education did not correlate significantly with any of the SIV scales for the Navy population. Rank correlated positively with Conformity ($r = .10$, $p < .05$) and Leadership ($r = .16$, $p < .05$). Intelligence (Navy General Classification Test) scores correlated significantly and in the expected direction with Conformity ($r = -.24$, $p < .001$) and Independence ($r = .10$, $p < .05$). It is noted that Gordon (1960) reported a significantly negative correlation between intelligence (The College Qualifications Test) and Conformity. It seems likely that the trends noted in Table 1 for the Conformity and Independence scales are not attributable to age or education; however, intelligence may account in part for the observed differences on those scales.

The pattern of intercorrelations among the SIV scales shown in Table 2 generally conforms to that reported by Gordon for a college student sample (1960, p. 5). Relationships among the scales, therefore, appear to be stable, being similar for these different populations. For both Navy and college populations the Support scale correlates most highly with other scales, being substantially correlated with Recognition and Leadership. The preponderance of negative correlations is due to the forced-choice method and resulting interdependence among scales.

Table 3 presents an analysis of relationships between 13 life status variables and the SIV measures for the total Navy sample, both officers and enlisted men (maximum $N = 497$). SIV score

TABLE 2

Intercorrelations among SIV Scales

	Conformity	Recognition	Independence	Benevolence	Leadership
Support	−26[a]	46	−14	−20	−51
Conformity		−33	−43	28	−32
Recognition			−29	−34	−10
Independence				−42	00
Benevolence					−28
Leadership					

[a] Decimal points are omitted.

distributions were dichotomized as closely to the median score as possible providing "high" and "low" groups on each test variable. Life status variables were grouped into categories and related to the dichotomized SIV measures by the χ^2 technique. Certain differences can be noted for the χ^2 results compared with those for the Pearson correlations described earlier for age, education, and rank versus the SIV scales, suggesting non-linearity of these relationships.

The Support scale was found to be significantly related to Rank. Officers scored lowest on this variable (35 percent above the median score) while unrated enlisted men scored highest (66 percent above the median). Age and years of naval service also were significantly related to Support, although the forms of these relationships were curvilinear. Men in the 30–32 age category and in the 9–11 years experience category, that is, persons falling near the middle of those distributions, tended to have lowest scores.

Conformity was significantly related to rank; the relationship was distinctly curvilinear with the middle of the rank distribution scoring high on Conformity and the extremes low.

Recognition was significantly related to Marital Status; the Separated-Divorced category accounted for most of the variance as 80 percent of this group scored above the median.

Independence scores were not significantly related to any of the life status variables at the .05 level of confidence.

Benevolence was related to rank, Deep Freeze experience, and frequency of worship. Benevolence scores tended to decrease with rank except for the first class petty officer group which scored relatively high. Men who had previously spent a year in the Antarctic tended to score low on the Benevolence scale (only 36 percent scored above the median). A distinctly linear relationship was

TABLE 3

Relationships Between Life Status Variables and Gordon Survey of Interpersonal Values Scales[a]

Variable	df	S	C	R	I	B	L
Rank	5	21.6[b]	11.9[b]	3.2	10.5	14.6[b]	14.8[b]
Age	7	15.9[b]	7.9	9.2	10.7	6.8	12.6
Years Service	6	15.7[b]	7.9	7.5	9.0	11.8	6.4
D.F. Experience	2	0.4	0.7	3.0	2.8	6.1[b]	1.7
Education	5	7.3	3.5	5.4	7.2	2.1	22.7[b]
Religion	6	9.1	9.6	5.8	12.4	5.6	6.4
Worship	3	3.8	6.2	5.5	4.0	8.1[b]	2.0
Number of Siblings	2	2.0	4.0	1.5	1.0	0.0	3.0
Family Mobility	6	2.6	5.3	7.1	1.8	2.2	4.4
Rural-Urban Residence	4	4.2	8.1	6.6	3.2	1.0	7.3
Marital Status	3	6.9	5.1	7.8[b]	2.0	4.3	8.5[b]
Parents' Marital Status	4	2.4	3.5	1.0	2.9	3.5	1.1
Region of Residence	8	8.9	3.3	4.4	8.1	4.3	7.8

[a] χ^2 statistics were computed from SIV scores dichotomized as near the median as possible versus life status variables grouped into varying numbers of categories; degrees of freedom (df) equal number of categories minus one.

[b] χ^2 significant beyond .05 level of probability.

present between frequency of worship and Benevolence; those who worshipped regularly scored high and those who never worshipped scored low.

The Leadership scale was significantly related to rank, education, and marital status. The relationship with rank was linear but discontinuous in that officers were much higher than enlisted men on this variable (76 percent above the median). Education was highly related to Leadership scores; individuals with college experience scored above the median much more frequently (77 percent) than other categories. The "eleventh grade completed" category scored lowest on Leadership. Separated or divorced individuals tended to score high on the Leadership measure (64 percent above the median).

The number of siblings (none, one, more than one), family mobility, rural-urban residence, parents' marital status, and region of childhood residence variables were not significantly related to any of the test variables.

Discussion

The findings indicated that values measured by the SIV inventory tend to be sensitive to differences in current social status and generally insensitive to differences in familial and cultural back-

ground (parents' marital status, number of siblings, region of childhood residence, rural-urban residence, family mobility, and religion.)

The specific positive findings from the study appear very reasonable in retrospect: that low rank and inexperience are associated with needs for support and encouragement is not surprising; that a college education should be correlated with willingness to exercise leadership seems predictable; that religious devotion should be associated with benevolent attitudes toward others seems congruent with expectations; and that there may be a higher probability of marital friction and break-up where husbands are highly concerned about receiving attention and admiration and at the same time with being dominant and completely in charge does not appear illogical. Such relationships have not been previously documented for the SIV; the findings provide some evidence for the concurrent validity of certain of the SIV scales.

Further support for some of the specific relationships observed in the present study is available from other sources. Gordon (1960) reported a correlation of .52 between the Religious Scale of the Allport-Vernon-Lindzey Study of Values and the SIV Benevolence Scale, indicating a close relationship between this measure of religious interest and benevolent attitudes. In a previous study, the authors (Gunderson and Nelson, 1964) noted that separated or divorced individuals scored significantly higher than other individuals on the Expressed Control Scale of the FIRO-B Inventory. Gordon (1963) reported a correlation of .60 between the SIV Leadership and the FIRO-B Expressed Control Scales for a sample of Antarctic scientists. These relationships tend to reinforce the plausibility of the common impression that particular syndromes of interpersonal needs or values in males may be inimical to successful marriage. The notion that regards sharp differences in interpersonal values as sources of marital incompatibility is not novel, but the specific clues suggested above may be worthy of further study.

The fact that men who had previously spent a year in the Antarctic tended to score lower (64 percent of the 44 subjects in this category scored below the median) on the Benevolence scale than men without such experience is not readily explained from available data. These men typically had experienced confinement and

enforced socialization for several months during the Antarctic winter and were again volunteering to spend a year living at close quarters. The possibility that benevolent attitudes undergo change in such an environment should be tested directly by obtaining measures on the same or comparable groups before and after Antarctic duty. It may be, however, that those actually selected for wintering-over assignments from among the many who apply score lower on the Benevolence scale or that a maturational process takes place such as that reported by Gordon (1963) for medical school students.

The results have demonstrated that the Survey of Interpersonal Values scales are significantly related to certain important aspects of current life status. Further research with the SIV should take these relationships into account.

Summary

Relationships of 13 life history and status variables to the six scales of the Survey of Interpersonal Values (SIV) were examined in a population of Navy volunteers for duty in Antarctica. The Support scale was significantly related to age, experience, and rank while Benevolence was related to rank, previous Antarctic experience, and participation in worship. These and other results suggested that the SIV value scales were sensitive to differences in current life status but generally insensitive to differences in familial and cultural background. Since the SIV was designed to measure present value orientations, the study provided evidence for concurrent validities of some of the scales.

REFERENCES

Gordon, L. V. *Manual for Survey of Interpersonal Values.* Chicago: Science Research Associates, Inc., 1960.

Gordon, L. V. *Research Briefs on Survey of Interpersonal Values.* Manual Supplement Revised. Chicago: Science Research Associates, Inc., 1963.

Gunderson, E. K. E. "Personal and Social Characteristics of Antarctic volunteers." *Journal of Social Psychology,* LXIV (1964), 325–332.

Gunderson, E. K. E. and Nelson, P. D. "Life Status Correlates of the FIRO-B Inventory in Navy Men." U. S. Navy Medical Neuropsychiatric Research Unit, San Diego, California, *Report No. 64–25,* 1964.

Nelson, P. D. and Gunderson, E. K. E. "Personal History Correlates of Performance among Military Personnel in Small Antarctic Stations." U. S. Navy Medical Neuropsychiatric Research Unit, San Diego, California, *Report No. 63–20,* 1963.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 1, 1966

CONTRASTED GROUPS VERSUS REPEATED MEASUREMENT DESIGNS IN THE EVALUATION OF SOCIAL DESIRABILITY SCALES[1]

GEORGE J. SKRZYPEK AND JERRY S. WIGGINS
University of Illinois

SOCIAL desirability scales are psychometric indices which purport to tell us the extent to which individuals' answers to personality tests are influenced by a tendency for the respondents to place themselves in a favorable light. Wiggins (1959) compared the relative success of several such scales in differentiating between subjects instructed to "fake good" and other subjects not so instructed on the MMPI. He found that some scales, such as his own and that of Cofer, Chance and Judson (1949) were able to separate such instructional groups with a modest amount of success while others, notably that of Edwards (1957) and the *K* scale (Meehl and Hathaway, 1946) were considerably less successful. Walker (1962) performed a partial replication of Wiggins' study and obtained results which reflected more favorably on Edwards' scale (*SD*) and the *K* scale as social desirability measures.

Since Walker (1962) employed different instructions and a test-retest design involving a single group of subjects it is not clear whether the discrepancies between his results and those of Wiggins should be attributed to the use of "improved" instructions as Walker (1962, 1963) maintains or to the use of a test-retest design as Wiggins (1963) asserts. Wiggins' preference for a contrasted groups design was stated as follows:

---

[1] This study is based on a Masters thesis submitted to the University of Illinois by the first author and conducted under the supervision of the second author. The study was supported by Public Health Service Research Grant No. MH 07042-01 from the National Institute of Mental Health.

In addition to providing a more realistic basis for generalization, the contrasted groups design avoids the necessity of employing a test-retest control group for the "improved adjustment" effect that frequently occurs with repeated administrations of the same inventory (Windle, 1954, 1955; Wiggins, 1963, p. 110).

Although both points of view in this controversy have been well explicated (Wiggins, 1959, 1962, 1963; Walker, 1962, 1963) the issues involved can be resolved only by additional experimentation. In the present study, improvement phenomena, contrasted groups versus test-retest differences, and the effects of different instructions were investigated within the context of a single experimental design.

Method

One hundred and fifty male undergraduate students enrolled in an introductory course in psychology were administered the full scale MMPI on two separate occasions separated by a one week interval. The subjects were divided into three groups of 50 each according to the experimental design illustrated in Table 1. Self-report instructions were those printed on the standard MMPI test booklet. Fake instructions were designed to approximate as closely as possible those employed by Walker (1962).[2]

TABLE 1

Experimental Design

N	First Administration	Interval	Second Administration
50	Ia Self-Report	one week	Ib Fake
50	IIa Fake	one week	IIb Self-Report
50	IIIa Self-Report	one week	IIIb Self-Report

MMPI protocols were scored for the standard clinical and validity scales as well as the six social desirability scales employed in the original study (Wiggins, 1959). These social desirability scales were: Edwards' *SD* (Edwards, 1957), Hanley's *Tt* (Hanley, 1957), Wiggins and Rumrill's *Sd-A* and *Sd-R* (Wiggins and Rumrill, 1959), Cofer's *Cof* (Cofer et al., 1949) and Wiggins' *Sd* (Wiggins, 1959). A brief description of the content and construction of these scales is given in Wiggins (1959). Scale means and standard deviations

[2] These instructions are given in Skrzypek (1964).

were computed for each group as a basis for t test comparisons among the groups. In addition, the proportion of hits and misses in group identification at different scale cutting scores was computed and the discriminative efficiency of each cutting score assessed by the phi coefficient.

The design presented in Table 1 allows for several comparisons whose outcomes may be anticipated on the basis of previous experimental findings and the interpretations that have been placed on them. When the scale scores of Group Ia are compared with those of Group IIa, the contrasted groups design of Wiggins (1959) is replicated. The comparison of Group Ia with Group Ib is the test-retest design employed by Walker (1962). A comparison of Group Ib with Group IIa involves a direct test of differences between test-retest and independent experimental (faking) groups. Such differences would be anticipated on the basis of the discrepant findings reported for Walker's (1962) and Wiggins (1959) experimental groups. Since both experimental groups had identical instructions, differences between them would be attributed to previous experience in self-report.

An additional order effect may be tested by comparing Group Ia with Group IIb. The results of Voas (1958) suggest that prior faking experience may tend to result in more "honest" subsequent self report.[3] In the earlier studies of Cofer et al. (1949) and Rosen (1956), no such order effects were obtained.

A comparison between Group IIIa and Group IIIb allows for an evaluation of a possible "improvement" effect attributable to simple retesting. According to the manner in which Wiggins (1963) interprets Windle's (1955) findings as applying to Walker's (1962) study, such an order effect would be evident in Edwards' *SD* scale.

Results and Discussion

Independent versus Retested Faking Groups

Table 2 contains the results of the replication of Wiggins (1959) study (Group Ia versus Group IIa), the replication of Walker's

---

[3] Although Voas (1958) required *Ss* to give self-report and fake answers at the *same* point in time, his suggestion that bringing the tendency to fake to a conscious level may inhibit faking in self-report would seem applicable to the present design as well.

(1962) study (Group Ia versus Group Ib) and a direct comparison of the experimental groups from both designs (Group Ib versus Group IIa). In contrast to Wiggins' original results, highly reliable differences between group means (Group Ia versus Group IIa) were obtained for *all* social desirability scales (including *SD*, *K* and *F*) in the present replication. Although the possibility of intrinsic differences in the subject samples employed in the two studies cannot be ruled out, Walker's (1962) contention that his version of faking instructions inspires greater faking behavior, receives support from the present findings. The replication of Walker's study (Group Ia versus Group Ib) provides results similar to those originally found, although the extent of mean differences in the *SD* scale is slightly less than that found by Walker (1962).

TABLE 2

Mean Scale Differences under Contrasted Groups and Test-Retest Designs

		L	F	K	SD	Tt	Sd-A	Sd-R	Cof	Sd
Contrasted Groups Comparison										
Group Ia		L	F	K	SD	Tt	Sd-A	Sd-R	Cof	Sd
Self-Report	$\bar{X}$	2.60	5.86	13.78	29.66	11.48	24.34	24.84	11.92	13.14
(N = 50)	σ	1.96	4.04	4.34	6.03	2.93	7.68	3.30	3.07	3.77
Group IIa	$\bar{X}$	9.30	3.96	20.02	36.30	18.68	33.88	29.18	23.90	27.36
Fake	σ	3.08	3.53	4.00	3.52	3.53	4.53	3.88	4.78	5.69
(N = 50)	t	12.85	2.48	7.40	6.66	11.00	7.49	5.97	14.77	14.58
	p	.001	.02	.001	.001	.001	.001	.001	.001	.001
Test-Retest Comparison										
Group Ia		L	F	K	SD	Tt	Sd-A	Sd-R	Cof	Sd
Self-Report	$\bar{X}$	2.60	5.86	13.78	29.66	11.48	24.34	24.84	11.92	13.14
(N = 50)	σ	1.96	4.04	4.34	6.03	2.93	7.68	3.30	3.07	3.77
Group Ib	$\bar{X}$	10.58	3.94	21.46	36.88	20.14	37.74	31.00	24.76	29.28
Fake	σ	2.84	3.60	3.09	2.54	2.28	3.24	2.43	4.51	4.93
(N = 50)	t	15.43	3.12	12.26	8.40	18.36	9.83	11.59	18.58	19.97
	p	.001	.01	.001	.001	.001	.001	.001	.001	.001
Contrasted vs Test-Retest Experimental Groups										
Group IIa vs	t	2.13	.03	1.99	.94	2.43	1.08	2.78	.92	1.78
Group Ib	p	.05		.05		.05		.01		

When scale scores from the two experimental groups (Group Ib and Group IIa) are compared, higher faking scores are found for all social desirability scales in the experimental group which had preceding self-report experience. These differences are reliable for only four scales (although *K* is one of them). The difference in *SD* scale scores for the two experimental groups is not a statistically reliable one. Nevertheless, it appears that faking groups with prior

self-report experience are not comparable to faking groups which have not had such experience.

The Effect of Prior Experience

The hypothesis that prior faking experience may lead to subsequently more "honest" self-report was tested by a comparison of mean scale scores between Group Ia (which had no prior experience) and Group IIb (which had prior faking experience). With one minor exception, prior faking experience resulted in *increases* in social desirability scale scores rather than decreases as predicted. None of these differences was statistically reliable however. When scores on the nine MMPI clinical scales were compared between these two groups, decreases were found to have occurred in all scales. The only clinical scale difference that attained significance was that of *Hs* ($p < .05$). If anything, prior faking experience leads to decreased "honesty" in subsequent self-report.

A comparison of Group IIIa and Group IIIb allows for an evaluation of a possible "improvement effect" which has been noted by Windle (1955) as occurring in test-retest self-report designs. In the present study, all of the social desirability scales *increased* on second administration in the absence of specific faking instructions. Edwards' *SD* scale was the only scale which increased reliably ($t = 2.25$; $p < .05$) adding a dramatic note to Wiggins' (1963) prediction of this finding. All of the clinical scales decreased ("improved") on retesting. Only *Hs* and *Hy* did so significantly ($p < .05$). Studies such as Walker's (1962) which evaluate the efficiency of Edwards' *SD* by means of a test-retest faking design may be capitalizing on an improvement phenomenon which operates in the direction of the experimental hypothesis.

Discriminative Efficiency

Although the emphasis thus far has been on mean scale differences between experimental and control groups, it should be clear that the ultimate basis for evaluating the usefulness of any screening device is the extent to which such a device contributes to the accuracy of personnel decisions (Cronbach and Gleser, 1957). This, in turn, may be seen to be a function of the amount of *overlap* which exists between experimental and control group score distributions and the proportion of correct and incorrect decisions

made at a given cutting score (Meehl and Rosen, 1955). In Table 3 are presented the proportion of correct and incorrect decisions made by each of the social desirability scales at the arbitrarily adopted cutting score at which both false positive and false negative decisions are minimized. The *relative* success of these scales in correctly identifying fakers and non-fakers has now been sufficiently replicated (Wiggins, 1959; Walker, 1962; Boe and Kogan, 1964) to assume the status of a minor psychometric law. The present findings with respect to independent experimental and control groups are highly similar to those of the original study (Wiggins, 1959) with the exception that all scales are more efficient, due perhaps to the use of Walker's instructions.

Although not presented, the findings with respect to discrimination between the same subjects taking the test under self-report and faking instructions are highly similar to those of Table 3.

TABLE 3

Discriminative Efficiency of Social Desirability Scales in Contrasted Groups Design

Scale	Cut	Group Ia Proportion called Self-Report	(Self-Report) Proportion called Fake	Group IIa Proportion called Self-Report	(Fake) Proportion called Fake	Phi Coeff.
Sd	22+	.98	.02	.10	.90	.883
Cof	18+	.96	.04	.10	.90	.862
L	7+	.96	.04	.16	.84	.806
Tt	16+	.90	.10	.16	.84	.741
Sd-A	32+	.78	.22	.12	.88	.663
Sd-R	30+	.98	.02	.42	.58	.611
K	18+	.80	.20	.20	.80	.600
SD	35+	.74	.26	.16	.84	.583

Hanley's *Tt* did relatively better under test-retest conditions and the overall degree of discrimination (as reflected in the phi coefficient) was greater for all scales, due perhaps to an improvement phenomenon. In both designs, *Sd* and *Cof* were among the three most efficient scales, while *K* and *SD* were the least efficient.

Summary

Wiggins (1959), using an independent groups design in evaluating several social desirability scales, found certain scales to perform

considerably more effectively than others. Using different instructions and a test-retest design, Walker (1962) obtained somewhat different results. The present study investigated the hypothesis that test-retest faking designs are subject to a spurious "improvement" effect attributable to simple retesting. It was found that the test-retest faking design results in higher faking scores than the design employing independent groups and this increment may, in part, be attributed to an "improvement" phenomenon.

REFERENCES

Boe, E. E. and Kogan, W. S. "Effect of Social Desirability Instructions on Several MMPI Measures of Social Desirability." *Journal of Consulting Psychology,* XXVIII (1964), 248–251.

Cofer, C. N., Chance, June, and Judson, A. J. "A Study of Malingering on the MMPI." *Journal of Psychology,* XXVII (1949), 491–499.

Cronbach, L. J. and Gleser, Goldine, C. *Psychological Tests and Personnel Decisions.* Urbana: University of Illinois Press, 1957.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden, 1957.

Hanley, C. "Deriving a Measure of Test-Taking Defensiveness." *Journal of Consulting Psychology,* XXI (1957), 391–397.

Meehl, P. E. and Hathaway, S. R. "The *K* Factor as a Suppressor Variable in the MMPI." *Journal of Applied Psychology,* XXX (1946), 525–564.

Meehl, P. E. and Rosen, A. "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores." *Psychological Bulletin,* LII (1955), 194–216.

Rosen, E. "Self-Appraisal, Personal Desirability and Perceived Social Desirability of Personality Traits." *Journal of Abnormal and Social Psychology,* LII (1956), 151–158.

Skrzypek, G. J. "Contrasted Groups Vs Repeated Measurement Designs in the Evaluation of Dissimulation Scales." Unpublished masters thesis, University of Illinois, Urbana, 1964.

Voas, R. B. "A Procedure for Reducing the Effect of Slanting Questionnaire Responses Toward Social Acceptability." Educational and Psychological Measurement, XVIII (1958), 337–345.

Walker, J. N. "An Examination of the Role of the Experimentally Determined Response Set in Evaluating Edwards' Social Desirability Scale." *Journal of Consulting Psychology,* XXVI (1962), 162–166.

Walker, J. N. "Social Desirability: A Reply to Wiggins." *Journal of Consulting Psychology,* XXVII (1963), 458.

Wiggins, J. S. "Interrelationships Among MMPI Measures of Dissimulation Under Standard and Social Desirability Instructions." *Journal of Consulting Psychology,* XXIII (1959), 419–427.

Wiggins, J. S. "Strategic, Method, and Stylistic Variance in the MMPI." *Psychological Bulletin,* LIX (1962), 222–242.

Wiggins, J. S. "Social Desirability Under Role-Playing Instructions: A Reply to Walker." *Journal of Consulting Psychology,* XXVII (1963) 107–111.

Wiggins, J. S. and Rumrill, C. "Social Desirability in the MMPI and Welsh's Factor Scales *A* and *R*." *Journal of Consulting Psychology,* XXIII (1959), 100–106.

Windle, C. "Test-Retest Effect on Personality Questionnaires." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIV (1954), 617–633.

Windle, C. "Further Studies of Test-Retest Effect on Personality Questionnaires." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XV (1955), 246–253.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

INHIBITORY EFFECTS OF A PRETEST ON OPINION CHANGE[1]

ROBERT E. LANA
University of Rome, Italy

In recently reported studies (Lana and Rosnow, 1963; Lana, 1964b) it has been shown that the use of a pretest and its manner of presentation can affect the direction of opinion change when two opposed arguments on the same topic are utilized. Primacy refers to the success in changing opinion of an initial argument of two opposed communications. Recency refers to a similar success of an argument presented second. When the pretest is hidden from the subject (Lana and Rosnow, 1963) or missing altogether (Lana, 1964b), in the pretest-treatment-posttest opinion change research design (Solomon, 1949), change occurs usually in the direction of the first presented of two opposed arguments (primacy). Recency, however, does not necessarily occur when the pretest is exposed.

The principal concern of this study is to examine the effect of the pretest on opinion change regardless of direction (primacy or recency). The initial focus of the studies mentioned above (Lana and Rosnow, 1963; Lana, 1964b) was on the effect of the present on the resulting directional effect (primacy-recency), but a more important consideration may be whether or not the pretest generally in-

[1] This research was supported by the National Institute of Mental Health, United States Public Health Service Grant No. MH 06926-02.

The Vivisection Questionnaire, the pro and con communications, and the familiarization talk utilized in this study have been deposited with the American Documentation Institute. Order Document No. 6765 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief; Photoduplication Service, Library of Congress.

hibits or facilitates opinion change when two opposed arguments on the same topic are presented to the subject. The concept here is that the subject may "commit" himself to a particular position with respect to the topic at hand. Thus responding to a pretest should result in less opinion change, either in favor of or against a specific issue, than where an initial pretest is not used. If this is true, then using a pretest in opinion and attitude studies will reduce the effect ordinarily produced by persuasion. Solomon (1949) discovered a depressive effect of the pretest on errors in spelling using school children as subjects. Entwisle (1961) found significant interaction effects with IQ and sex in a similar training situation. Both utilized a single, unidirectional treatment.

Hovland and Mandell (1957) have shown that commitment to an attitudinal position inhibits change when commitment is elicited after an initial influence attempt but not when response to a precommunication questionnaire constitutes the commitment. Their suggestion, however, is that there are forms of attitudinal commitment, usually made under public conditions, which inhibit opinion or attitude change as a result of materials presented later.

A subject's initial response to a questionnaire may provide a basis of comparison for further questionnaire responses and thus a positive correlation of some magnitude might be expected between the individual's first and second responses to the same questionnaire. This expectation of correlation between successive responses to some task by the same subject is the basis for methodological concern with repeated measurements in experimental design. Statistical psychologists have been concerned with the problem for years, and there seems to be little doubt about the influence of one response on another in the repeated measurement situation.

When two treatments are performed in succession, a treatment carryover may occur. The rotation design (Cochran and Cox, 1957) is specifically intended to give information on such a treatment carryover. In the experimental situation described in this paper where a pretest precedes a single treatment, the possible effects of carryover from pretest to treatment are the same as the treatment to treatment carryover. A rotation design is not possible when the first factor is a pretest since, by definition, it must precede the treatment. Consequently any examination of the confounding effects of the pretest with treatment must be made by experimentally

manipulating the nature of the pretest. The major question that remains concerns the nature of the relationship. At any rate, there seems to be ample information to establish the suspicion that initial responses of a subject in the opinion and attitude situation may very well depress later changes.

In a study done in 1959, Lana found no depressive effect of the pretest (interaction between pretest and treatment) when a single, unidirectional communication was utilized. However, when two opposed communications on the same topic were presented to the same subject (Lana and Rosnow, 1963, 1964b) a pretest inhibition or depressive effect was present. Recalculation of the data of these studies was necessary, and the depressive effects were not reported at the time of publication.

In order to facilitate comparisons among the various experiments already reported in the literature and the current one, the procedure, results, and conclusions of each of the three previously published experiments are listed below. All *primacy-recency effects* are summed by taking the algebraic differences between pretest and posttest and subtracting these differences for each pro-con and con-pro group. These results have already been reported in the first three published studies cited below and, for the fourth study, in Table 2 of this report. All *opinion change effects regardless of direction* (primacy or recency), are absolute differences, taken between pre and posttest for all groups and averaged regardless of algebraic sign.

Experiment I (Lana 1959) Exposed Pretest, Unidirectional Persuasion.

One-hundred fifty-six college students were randomly assigned to four treatment conditions. Two of these groups received the pretest attitude questionnaire. One of the two groups listened to a taped pro-vivisection communication 12 days after taking the pretest. After treatment, this group (Group I) was immediately posttested with the same questionnaire. The other group (Group IV) was simply posttested 12 days later. Group II heard the pro-vivisection communication and was posttested immediately afterward without having been pretested. Group III answered the questionnaire once. Opinion change was measured as the algebraic difference from pretest to posttest.

The only significant F was that for the treatment effect, while the F ratios for the effects of the pretest and the pretest-by-treatment interaction were both less than unity. It can be concluded that the pretest and treatment did not interact, and that the pretest did not sensitize or desensitize the subjects to the communication. The main effect of pretesting was not significant.

Experiment II (Lana and Rosnow 1963) Exposed Versus Hidden Pretest, Bidirectional Persuasion

One-hundred twenty-eight students from beginning level psychology classes at the American University were used as subjects in this study. They were randomly divided into eight groups of 16 each. All S's initially received either the opinion questionnaire on nuclear weapons or one on public censorship. Four of the groups were presented the pretest during a regularly scheduled class as a separate part of the hour's activity. These were labeled the "exposed pretest" group. The remaining four groups responded to the questionnaire presented as part of their regularly scheduled hour examination in general psychology. The opinion questions appeared as a 5-question unit in the exam and were "buried" among other similar question units pertaining to the usual subject matter of an introductory psychology class. These groups were labeled the "hidden pretest" group. All groups, immediately after listening to the pro and con arguments, received the appropriate posttest, which was identical with the pretest.

Primacy-recency effects are given in the 1963 report on this experiment. Recalculation of the data indicated that the average opinion change per group (mean absolute differences from pretest to postest) of the four groups where the pretest was hidden was 1.88 scale points as compared with an average absolute change of .41 scale points per group of those four groups where the pretest was exposed. This difference was significant beyond the .05 level (t-test for independent means).

Experiment III (Lana 1964b) Exposed versus No Pretest, Bidirectional Persuasion

One-hundred thirty-six subjects from the introductory psychology classes at Alfred University were randomly assigned to 12 experimental groups. Two of the groups were "zero" control groups

which received the pretest and then the posttest either 12 or 24 days later. Neither of these groups indicated any change in opinion whatsoever and thus were dropped from any further analysis. Of the remaining 10 groups, eight were pretested with an opinion questionnaire on Nikita Khrushchev. Pretests for all groups were then analyzed for differences with a one-way analysis of variance to determine whether or not the groups could reasonably be assumed to be homogeneous with respect to initial opinion concerning Khrushchev. The F ratio was less than 1. It was concluded that the groups were homogeneous on initial opinion. It was assumed that initial opinions of the two groups not pretested were homogeneous with the eight pretested groups since they were formed randomly from the same sample from which these latter groups were selected.

Four of these eight groups, immediately after filling out the pretest, were exposed to the successive presentation of two opposed (pro and con) prose statements regarding Khrushchev. Two of these groups received the pro argument first and the other two groups the con argument first. Two of these four groups were then immediately posttested. The remaining six groups, and the two groups that were not initially pretested, were dismissed for 12 days. At the termination of this interval, all eight groups who had not as yet been given a posttest were run as follows: The two groups that had been exposed to the pro and con arguments were posttested with the identical questionnaire used as the pretest. The two groups receiving no pretest were given the pro-con or con-pro order of arguments and immediately posttested. Two of the groups having received the pretest only received the pro-con or con-pro order of arguments and were then posttested. The remaining two groups, who had received the pretest only, were exposed to the arguments as had the other groups. They were then posttested 12 days after the presentation of the arguments. Because this design is more complicated than those of the other reported studies, a summary of the experimental procedure is given in Table 1.

In the two groups receiving no pretest, the average opinion change (mean absolute difference from pretest to posttest) per group was .75 scale points as compared with an average per group change of .35 scale points for the groups receiving a pretest. These differences are significant beyond the .05 level (t-test for independent means).

TABLE 1

Experimental Design for Experiment III

GROUPS											
A	B	C	D	E	F	G	H	I	J	K	L
Pretest		Pretest				Pretest		Pretest		Pretest	
Pro	Con	Pro	Con	Pro	Con	12 days		12 days		12	24
Con	Pro	Con	Pro	Con	Pro	Pro	Con	Pro	Con	days	
12 days		Posttest$_1$		Posttest$_1$		Con	Pro	Con	Pro		
Posttest$_1$						12 days		Posttest$_1$		Posttest$_1$	
						Posttest$_1$					
3-months delay											
Posttest$_2$											

Thus, in Experiment I, with a unidirectional influence attempt, no damping of the posttest by the pretest was found. In Experiment II, with a hidden pretest and bidirectional persuasion, there was no damping effect of pretest on posttest, but there was a damping effect when the pretest was exposed. In Experiment III with bidirectional persuasion and no pretest, there was no damping effect (estimated); with an exposed pretest, there was a damping effect. For procedural details and analysis of primacy-recency effects the original publications should be examined.

Experiment IV—Current Study

The primary hypothesis is that a pretest will dampen the effect of bidirectional persuasion. A second hypothesis was also tested in this study. Previous research has shown that a subject familiar with the topic of a communication is more likely to change his opinion in the direction of the first presented of two opposed arguments (primacy effect) when that topic is concerned with some social issue. As stated above, it also seems that a primacy effect is prevalent in cases where the subject is not pretested. In combining these two conditions, high familiarity and no pretest, a pronounced primacy effect should occur.

One-hundred subjects in the introductory psychology classes at Alfred University were randomly assigned to one of eight experimental groups. All subjects were asked to answer two questions on a sheet of paper. They were: "If you know what the word, 'vivisection' refers to, describe it in one of two sentences," and "Also describe, in one or two sentences, any recent events that you know of concerning

vivisection." Only about ten percent of all subjects knew to what vivisection referred. This group, and randomly chosen subjects from the unfamiliar students, were designated the "familiar group." The remainder of the subjects were designated the "unfamiliar group." Just prior to pretesting, and again just prior to posttesting, the familiar group was read a 213 word passage describing vivisection. The information presented was factual and did not represent any attitudinal point of view. Thus the familiar group, some of the members of which already possessed some knowledge of the topic, was exposed to information about the nature of vivisection. The familiar and unfamiliar groups were randomly divided, half receiving a questionnaire tapping opinions for or against the practice of vivisection. These groups were again randomly subdivided into groups receiving a pro-vivisection communication followed by a con-vivisection communication, or the con-communication and then the pro-communication. The scores on the pretest for the four groups were analyzed for differences by a one-way analysis of variance to determine whether the groups could reasonably be assumed to be homogeneous with respect to initial opinion on vivisection. The F ratio was less than one. It is concluded that the groups were homogeneous on initial opinion. It is assumed that the four remaining groups not pretested are homogeneous with the four pretested groups since they were formed randomly from the same sample as the pretested groups. In a single study such as this one, assumptions of this nature might seem tenuous. However, in other studies (Lana, 1959, 1964b) there seemed to be reasonable assurance that one could assume that unpretested groups had similar initial opinions as pretested groups. Random allocation assured homogeneity of opinion for the variously formed experimental groups over the long run. In any case, random allocation is one of the few techniques available at this time which is useful in creating the conditions necessary for making a test of the current hypothesis. One alternative to administering a pretest, and being reasonably certain of the opinion of the group on some topic, is to use groups which have been formed because of their stand on the very topic in question. Other difficulties, however, arise with this technique (Lana, 1964b).

All groups were then posttested with the identical questionnaire used as the pretest. The design is presented in Table 2.

Materials

A Likert-type, 19-item, 6-alternative questionnaire (Molnar, 1955) was utilized as both the pretest and posttest. It has been used in two previous studies by Lana (1959, 1961). High score indicates pro-vivisection opinion. The range of possible scores is 6 to 60. The pro and con communications on vivisection are about 500 words long and take about five minutes to read. They were developed by Molnar (1955) and have succeeded in changing the opinions of college students toward vivisection in other similar studies (Lana, 1959, 1961; Molnar, 1955).

Vivisection was chosen as the topic since it is generally not considered particularly controversial (Lana 1962) by most college students. Also, from the results of the initial questions asked, many students have no idea as to what the term refers. This makes it easy to familiarize the appropriate groups with the topic by providing vivisection information to them, thus increasing the probability that the familiar and unfamiliar conditions have been established in the experimental groups. Zero control groups (no communication) were not included in the design because in all similar instances (Lana, 1959, 1964b) no pretest-posttest changes were evident.

Results

The results of the experiment were analyzed by the use of the t-test for significance of average absolute differences between pre and posttest and the subtractive-difference technique, which is useful for detecting directional effects in opinion change. The groups not pretested were assigned mean pretest values based on the average of the scores of the pretested subjects. The average absolute opinion change score per unpretested group was 2.18 scale points regardless of direction (pro or con). Average absolute opinion change in the pretested groups was .95 scale points. The difference between the two groups is significant beyond the .01 level. The unpretested groups shifted 28 percent of the total distance it was possible for them to change on the opinion questionnaire continuum. The pretested groups shifted 12 percent of the maximum amount that it was possible to change.

Significant (.01) recency effects were evident in the familiar groups, and no significant directional effects of any kind resulted

in the unfamiliar groups. There were no differences in significant primacy-recency effects between the pretested and unpretested groups. The results are summarized in Table 2.

TABLE 2

Experimental Design and Summary of Primacy—Recency Results

	2 Questions on Vivisection—3 Days							
	FAMILIAR INFORMATION ON VIVISECTION				UNFAMILIAR NO INFORMATION ON VIVISECTION			
	NO PRETEST		PRETEST		NO PRETEST		PRETEST	
	12 DAYS							
	PRO CON	CON PRO	PRO CON	CON PRO	PRO CON	CON PRO	PRO CON	CON PRO
	INFORMATION ON VIVISECTION				NO INFORMATION ON VIVISECTION			
	Posttest							
	(EST.)	(EST.)			(EST.)	(EST.)		
PRE	37.43	37.43	37.00	35.71	37.43	37.43	39.92	37.36
POST	36.18	39.29	36.45	38.07	40.00	40.45	39.33	37.07
GAIN	−1.25	+1.86	−.55	+2.36	+2.57	+3.02	−.59	−.29
DIFF. OF DIFFERENCES	−3.11 (recency)		−2.91 (recency)		−.45 (recency)		−.30 (recency)	
n	11	14	11	14	13	11	12	14
t	**3.99		**4.04		<1		<1	

** p < .01

Discussion

The results indicate that the first hypothesis has been supported. Opinion change was significantly greater under conditions where no pretest was utilized than under conditions of pretesting. Unpretested, disguised pretest, and pretest groups (Lana, 1964b; Lana and Rosnow, 1963), and groups in preliminary experiments to the current study, were examined for differential amounts of opinion change. In all cases the no-pretest and disguised pretest groups changed to a greater degree than the pretested groups. This represents observations on 520 subjects. There seems to be some support for the contention that the pretest can act as a device by which the individual commits himself to maintain his opinion in the face of opposed (bidirectional) arguments presented later. Presumably, a single, unidirectional communication allows for greater change

regardless of initial pretest conditions because, if the individual is initially in favor of the position advocated by the communication his initial commitment is supported. If about half of all subjects are in favor of a given argument (assuming a normal distribution of pretest scores concerning the issue at hand, an assumption which holds true for this study) this half need not consider their initial commitment. On the other hand, when two opposed arguments are presented, there will always be one position confronting every subject which is essentially opposed to his own initial commitment. Thus, his commitment via pretest may influence his reaction to material presented later in the experiment, and this will be true for all subjects.

The pretest seems most effective in minimizing opinion change when the subject is later confronted with two opposed arguments on the topic in question rather than when he is later presented with a single, unidirectional argument. Conceivably one may speculate that the more complex and multi-sided the communicative materials, the greater the inhibitory effect of the pretest. Also it may be possible that the length, complexity and obviousness of the intent of the pretest might produce less opinion change after a communication than when a pretest is relatively simple, short and disguised.

The secondary hypothesis of the study, that the combined effects of familiarity of the subject with the topic, and the lack of a pretest, will produce a strong primacy effect was rejected. Precisely the opposite occurred. A strong recency effect was evident for the high familiar groups and no significant order effect of any kind resulted in the groups who were relatively unfamiliar with the topic. Recency effects were predicted for these latter groups. There were no differences in order effects between the pretested and the unpretested groups. In previous studies, primacy had been produced by high familiarity of the subjects with the topic and when the pretest was hidden or absent. This reversal of the predicted effects for the familiar group is surprising and no ex-post-facto interpretation of these results is entirely satisfactory. However, certain contingencies may be forwarded as possible sources of explanation.

The topic of vivisection has characteristically been rated as one of low controversy by college students. It has been shown (Lana, 1963a; 1964a) that topics of low controversy tend to produce no significant primacy-recency effects. This might have mitigated

against the appearance of primacy throughout all the groups of the study. It is also obvious that the familiarity of the subjects was more important in producing order effects than the fact of having been, pretested or not. In the current experiment, unlike previous studies on familiarity (Lana, 1961; Rosnow and Lana, 1965), the familiarization talk was presented twice, once before the pretest and again before the communications. This, and the questions used to measure existing familiarity in the subjects towards vivisection, probably provide a better assurance of familiarity for the subjects of this study than for those of previous studies. Even with these differences in methodology between the present study and the others, there is no convincing reason why recency should have dominated under conditions of familiarity.

Summary and Conclusions

One-hundred subjects were exposed to various experimental conditions where familiarity-unfamiliarity and pretesting-no pretesting were manipulated to examine their influence on magnitude of, and order effects in, opinion change. The following conclusions were made:

1. The taking of a pretest in the form of an opinion questionnaire inhibits opinion change regardless of the direction of change (primacy-recency) that may be strongest when opposed arguments are utilized.
2. Familiarity of the subjects with the topic of the influence attempt does not necessarily result in a significant primacy effect as predicted.

Theoretical interpretations of primacy-recency opinion change and, perhaps, measurement methodology, need to be revised and sensitized to establish consistent, successful predictions of these effects.

REFERENCES

Cochran, W. G. and Cox, G. M. Experimental Design (2nd ed.). New York: Wiley, 1957.

Entwisle, D. R. "Interactive Effects of Pretesting." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXI (1961), 607–620.

Hovland, C. I. and Mandell, W. "Is There a 'Law of Primacy' in Persuasion." In C. I. Hovland, et al, *The Order of Presentation in Persuasion.* New Haven: Yale University Press, 1957.

Lana, R. E. "Pretest-Treatment Interaction Effects in Attitudinal Studies." *Psychological Bulletin,* LVI (1959), 293–300.

Lana, R. E. "Familiarity and the Order of Presentation of Persuasive Communications." *Journal of Abnormal and Social Psychology,* LXII (1961), 573–577.

Lana, R. E. "Order Effects in Persuasive Communications." Progress Report to the National Institute of Mental Health, United States Public Health Service, Research Grant M4830, May 1962.

Lana, R. E. "Controversy of the Topic and the Order of Presentation in Persuasive Communications." *Psychological Reports,* XII (1963), 163–170. (a)

Lana, R. E. "Interest, Media, and the Order Effects in Persuasive Communications." *Journal of Psychology,* LVI (1963), 9–13. (b)

Lana, R. E. "Three Theoretical Interpretations of Order Effects in Persuasive Communications." *Psychological Bulletin,* LXI (1964), 314–320. (a)

Lana, R. E. "The Influence of the Pretest on Order Effects in Persuasive Communications." *Journal of Abnormal and Social Psychology,* LXIX (1964), 337–341. (b)

Lana, R. E. "Order Effects in Persuasive Communications." Progress Report to the National Institute of Mental Health, United States Public Health Service, Research Grant MH 06926, May 1964. (c)

Lana, R. E. and Rosnow, R. L. "Subject Awareness and Order Effects in Persuasive Communications." *Psychological Reports,* XII (1963), 523–529.

Molnar, A. "The Effects of Styles, Speeches, and Arguments on the Attitudes and Perceptions of a Listening Audience." Unpublished master's thesis, University of Maryland, 1955.

Rosnow, R. L. and Lana, R. E. "Complementary and Competing Order Effects in Opinion Change." *Journal of Social Psychology,* LXVI (1965), 201–207.

Solomon, R. L. "An Extension of Control Group Design." *Psychological Bulletin,* XLVI (1949), 137–150.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

THE SEARCH FOR RECURRING PATTERNS AMONG INDIVIDUAL PROFILES[1]

WILSON H. GUERTIN
University of Florida

In clinical medicine the symptoms of a patient are matched to a disease syndrome to establish the initial diagnosis. The clinical psychologist for many years has tried to find test patterns which will have diagnostic value so that he, too, can match the patient's test profile (symptoms) to a diagnostic model. While the approaches utilize the same method, the practical aspects of the tasks differ.

In clinical medicine, each disease usually has a single syndrome picture, occasionally with variants, so that a stable, useful model can be established and applied uniformly in practice. If an internist states that a patient showed the usual symptoms (syndrome) of obstructive jaundice, another would know what symptoms were implied. When a psychiatrist states that his patient showed the usual symptoms of schizophrenia, another psychiatrist would know only that the patient was psychiatrically ill and would be uncertain which symptoms might be present and which absent.

As psychologists, we must assume that a number of different models will be required to describe the members of a single diagnostic group such as those having "schizophrenia." The notion of isomorphic correspondence between psychiatric classification and symptom pictures may be a convenient simplification, but it is unrealistic.

Since Wechsler intelligence scales have been prominent in the

[1] The author is indebted to Drs. Douglas E. Scates, Ralph M. Reitan, Clayton E. Ladd, and Frank M. du Mas for their critical reading of the manuscript. Dr. Scates was particularly helpful in clarifying the term, mean profile.

search for diagnostic patterns, it is appropriate to turn to Wechsler's earliest proposals of patterns. Wechsler (1944) proposed a single pattern for each psychiatric diagnosis. During the subsequent 15 years he has not acknowledged in his manuals the need to propose multiple patterns for any of the diagnostic groups (Wechsler, 1958).

Those who try to test the validity of Wechsler's proposed patterns fail to find any description of how they were derived. It is usually assumed, since there is only one pattern for each diagnosis, that the pattern is a profile of the mean subtest scores for a sample of patients from that diagnostic group.[2] But the profiles of the mean subtest scores of actual samples of patients are almost certain to show much less variance (scatter) than found in Wechsler's patterns. Possibly the patterns proposed are those that he felt were representative of *S*s most typical of each diagnostic group. If this is so, Wechsler has presented modal patterns[3] rather than profiles of the means for actual groups.

Modal profile means a hypothetical profile of test scores that is typical of a number of individuals in a diagnostic group. It is the average profile obtained from a tight cluster of individuals with congruent profiles. The term *modal* refers to the frequent recurrence of the profile for members of the diagnostic sample. It should not be supposed that the statistic, mode, enters its derivation. In a computational sense the modal pattern *is* a profile of mean test scores in a battery. However, the mean scores are computed for a modal group (cluster) of similar profiles with the result that mean scores will not be intermediate values that seldom or never occur for actual subjects.

It is held that only modal patterns are suitable for use as diagnostic models. A modal pattern is a close approximation to a number of profiles actually obtained by *S*s. The usual profile of the means,

---

[2] Designation of *profile of the mean* test scores as *mean profile* of test scores leads to confusion in thinking. Psychological literature often makes reference to the "mean profile" of Wechsler subtest scores as if somehow the *profiles* of the individual *S*s had been averaged.

[3] Throughout the paper an arbitrary, but useful, distinction is made between *profile* and *pattern*. Profile is used to designate any set of test scores derived from an instrument (s). Profile is therefore the generic term whereas pattern is used to refer to a particular recurring profile that may have diagnostic significance. The pattern is the diagnostic model to which an individual's profile is matched.

on the other hand, is a whole-group statistic and every individual in the group can have a profile *quite dissimilar* to the profile of the means. It is desirable that we concentrate on the development of modal patterns, which may typify individuals, since diagnosis deals with individuals, not with groups. Profiles of the means from diverse subjects have little value in diagnostic investigations.

Wechsler probably failed to recognize clearly that his proposed patterns were intended to serve as modal patterns rather than profiles of the means or he would have seen the need to present more than one pattern for each group. Or, he may have hoped that others would work in this badly understood area to develop multiple modal patterns for each diagnostic group from quantified data. But there has been almost no progress in the development of such multiple modal patterns; in fact, modal patterns have failed to receive consideration, while study after study presents analyses in terms of the profiles of the means. The absence of quantitative techniques for analysis of profiles is sufficient to account for the neglect of modal patterns by investigators.

Except for the procedure suggested by Saunders and Schueman (1962), useful techniques of analyzing out modal patterns from masses of data have not been forthcoming. Considerable progress has been made in our thinking about the very fundamental question as to what is meant by similarity in profiles (Cattell, 1949; Cronbach and Gleser, 1953; du Mas, 1949; Haggard, *et al.*, 1959). Without such clarification in our thinking, we cannot proceed to find modal patterns which depend upon clusters of *similar* profiles for derivation.

Should profiles be similar in shape, or level, or variance (scatter)? or any two of these? or all three? There is no simple answer because the importance of any of these three features is a function of the particular diagnostic circumstances. For example, it is unrealistic to expect all hysterics to have the same overall native intellectual endowment, and if a *S*'s Wechsler profile fits the variance and shape of the pattern but not the level, we could not conclude that he was not a hysteric. On the other hand, perhaps a type of left-hemisphere brain lesion will lower the level of functioning to a fairly constant level for all such cases despite original level; then level might prove to be a very important differentia for this modal pattern. Each clinical category studied requires a re-evaluation of

how the similarity criterion should be established, but all three features of the profiles must receive some initial consideration.

The usual subtest scatter scores (deviation of each subtest score from the individual's average level of functioning) provide a profile that retains shape and variance but ignores level. Product-moment correlation of two profiles of subtest scores also ignores level but also disregards differences in average variance (amount of scatter) in the two profiles. Rank-difference correlation coefficients completely ignore differences in variance (as well as level) in subtest scores so that the profile of a normal person (low scatter) can correlate very highly with a badly deteriorated organic patient (large subtest scatter). For example, the profile of mean Wechsler subtest scores of Rappaport's normal control group correlated .90 (rank-difference method) with the profile for a group of deteriorated, diffuse brain-damaged patients (Matthews, Guertin, and Reitan, 1962).

Some statistics have been proposed that attempt to consider shape, level and variance in arbitrarily weighted degrees when comparing two profiles. While shape is a prime criterion for pattern similarity for most diagnostic groups, it is unlikely that correlational techniques alone will identify diagnostic modal patterns.

Illustration

The general problem of finding clusters of similar profiles in a diagnostic sample of people is reducible to two phases. First, is the mathematical definition of similarity. Secondly, these indices of similarity are examined through matrix algebra to establish bounds on the clusters of similar profiles.

There is some similarity between searching for modal patterns among many individual profiles and the examination of a large intercorrelation matrix for clusters. In both cases, there are the problems of what is related to which others, and when is a cluster an independent cluster rather than a subcluster.

Similarity usually means that two or more profiles will be reasonably congruent when graphed on the same coordinates. That is, they must be comparable in all respects. In Figure 1 we could say that profiles 1, 2, and 3 are similar to one another. The same is true for profiles 7, 8, and 9; and for profiles 4, 5, 6; as well as for profiles 10, 11, and 12.

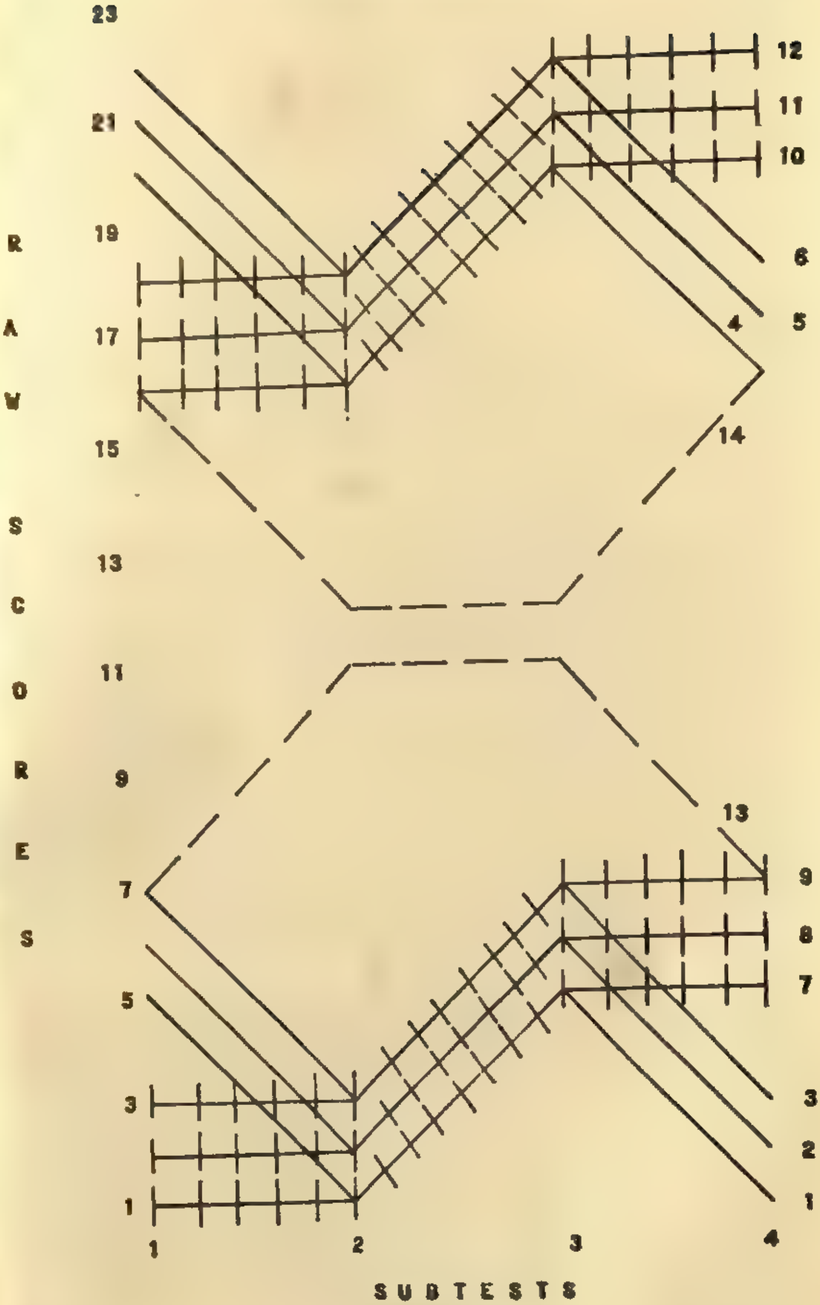

RAW SCORES
23
21
19
17
15
13
11
9
7
5
3
1
12
11
10
6
5
4
14
13
9
8
7
3
2
1
1
2
3
4
SUBTESTS

Figure 1.

Product-moment correlation coefficients depict similarity of profiles with respect to shape only. Thus, in Figure 1, profiles 1, 2, and 3, will be perfectly correlated with one another as well as with profiles 4, 5, and 6. Yet it is obvious that while the six profiles are similar in shape, they are not congruent because they are split into two markedly different levels.

The class of distance measures, such as Mahalanobis' *D*-Square, (1936), reflect similarity with respect to level and variance as well as shape. These distance measures are indices of the difference in heights of two profiles at each of the subtest score points. Distance measures come closer to representing what is meant by saying that two profiles are similar. Yet when a *D*-Square for two profiles is large it does not give any hint as to why the profiles are dissimilar. A large *D*-Square may arise from differences in shape, level, or variance, and, of course, any combination of these.

One unfortunate characteristic of the *D*-Square index is that two profiles with opposite shapes will yield a small *D*-Square when they are at the same levels and variance is small. If, as experience suggests, shape is of prime importance in classifying people from their profiles, the sole dependence on *D*-Squares will obscure some modal patterns.

Most psychologists are familiar with the factor analysis of correlations between people on a number of variables or test scores. Such an analysis is variously called transposed, inverted or *Q* factor analysis. Stated in terms of the pattern problem, the factoring of inter-profile correlations produces factors (clusters) of similarly-shaped profiles. Level and variance differences are, of course, ignored.

Factor analyses of distance measures are rarely reported in the literature. But enough is known to expect them to produce clusters of profiles similar with respect to level and variance, and usually with respect to shape as well. (Nunnally, 1962; Overall, 1964)

The procedure to be described herein, first factor analyzes inter-profile correlation coefficients. Then, the similarly-shaped profiles are examined by *D*-Square comparisons to see if they split into two or more clusters on the basis of levels and/or variance differences.

For example, in Figure 1, the perfectly correlated profiles 1, 2, 3, 4, 5, and 6 should split into two clusters at different levels when the inter-profile *D*-Squares for the six profiles are factor analyzed.

The same is true for profiles 7, 8, 9, 10, 11, and 12. The procedure now to be described is eminently successful in doing this.

A completely automatic computer package has been prepared by the author to do the successive computations described and illustrated here. Output from the computer is lengthy, but the essential results will be presented below from the analysis of the dummy problem represented by the profiles in Figure 1.

Two distinct shapes of profiles are seen in Figure 1. They are superimposed upon one another at the upper and at the lower level. The intercorrelation matrix, as seen in Table 1, contains only zeroes and ones:

The reduced intercorrelation matrix is factor analyzed by the Principal Axis method and rotated to the Varimax criterion to give the factor matrix in Table 2.

TABLE 1

Reduced Intercorrelation Matrix with Communalities in the Diagonals

	1	2	3	4	5	6	7
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.
7	0.	0.	0.	0.	0.	0.	1.0000
8	0.	0.	0.	0.	0.	0.	1.0000
9	0.	0.	0.	0.	0.	0.	1.0000
10	0.	0.	0.	0.	0.	0.	1.0000
11	0.	0.	0.	0.	0.	0.	1.0000
12	0.	0.	0.	0.	0.	0.	1.0000
13	0.	0.	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.	0.	0.

	8	9	10	11	12	13	14
1	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.
4	0.	0.	0.	0.	0.	0.	0.
5	0.	0.	0.	0.	0.	0.	0.
6	0.	0.	0.	0.	0.	0.	0.
7	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
8	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
9	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
10	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
12	1.0000	1.0000	1.0000	1.0000	1.0000	0.	0.
13	0.	0.	0.	0.	0.	1.0000	−1.0000
14	0.	0.	0.	0.	0.	−1.0000	1.0000

TABLE 2

Rotated Factor Matrix from the Intercorrelated Profiles

Profiles	Rotated Factor Loadings I	II	III
1	1.0000	0.	0.
2	1.0000	0.	0.
3	1.0000	0.	0.
4	1.0000	0.	0.
5	1.0000	0.	0.
6	1.0000	0.	0.
7	0.	1.0000	0.
8	0.	1.0000	0.
9	0.	1.0000	0.
10	0.	1.0000	0.
11	0.	1.0000	0.
12	0.	1.0000	0.
13	0.	0.	−1.0000
14	0.	0.	1.0000

The next stage is to analyze separately each of the factor columns in the rotated matrix to identify which profiles have high loadings on each of the factors. Factor columns are dropped unless three profiles have sizeable loadings (.40). This stage controls all future analyses because it feeds the information about loaded profiles forward for *D*-Square computation, factor analysis and computation of modal patterns. Each shape factor is completely analyzed through to the identification of modal patterns before it comes back here for the processing of the next shape factor.

TABLE 3

Transformed D-Squares among Profiles That are Loaded on the First Shape-Family Factor

Profiles	Profiles 1	2	3	4	5	6
1	288.0000	288.0000	285.0000	64.0000	33.0000	0.
2	288.0000	288.0000	288.0000	93.0000	64.0000	33.0000
3	285.0000	288.0000	288.0000	120.0000	93.0000	64.0000
4	64.0000	93.0000	120.0000	288.0000	288.0000	285.0000
5	33.0000	64.0000	93.0000	288.0000	288.0000	288.0000
6	0.	33.0000	64.0000	285.0000	288.0000	288.0000

The next stage is to compute *D*-Squares by the formula:

$$\frac{\sum_{j=1}^{L} (X_{ij} - X_{mj})^2}{L}$$

Where X is the raw score on the subtest j; and i and m are the two profiles being compared; and L is the number of subtest scores.

The D-Squares are reversed in direction by subtracting each from the largest obtained D-Square (289.00). After transformation, the largest raw D-Square will have a value of zero. The largest transformed D-Square of each column serves as the estimate of communality and is inserted in the diagonals in Table 3.

Next is the factoring and rotation of the transformed inter-profile D-Squares. The rotated factor matrix is shown in Table 4. Here we see that the loadings on the D-Square factors indicate the presence of two patterns—one for profiles 4, 5, and 6; and the other for profiles 1, 2, and 3.

TABLE 4

Rotated Factor Matrix Obtained from Analyzing Transformed Inter-Profile D-Squares for Profiles That are Loaded on the first Shape-Family Factor

	Factor	
Profile	I	II
1	0.0474	17.0355
2	1.8951	16.8841
3	3.6370	16.6268
4	16.6273	3.6378
5	16.8842	1.8951
6	17.0353	0.0470

Keep in mind we are still analyzing only the first shape family of profiles, corresponding to the first column of the rotated factor matrix (Figure 2) derived from the intercorrelations of the raw profiles. Now, the D-Square factor matrix is searched for profiles loaded heavily on the factors. Two profiles must be loaded heavily, or three moderately to establish a modal pattern. Pivot tests in the first column of the factor matrix are identified and then the next stage computes the modal pattern.

This next stage computes the scores which describe the modal pattern. The values are the mean scores of the pivot profiles (numbers 4, 5, 6) on the variables, weighted in relation to the size of the D-Square factor loadings of each of the pivot profiles.

The four obtained score values for the first modal pattern are: 21.02, 17.02, 21.02, and 17.02. These correspond closely to the true values of Figure 1, that are: 21.00, 17.00, 21.00, and 17.00.

After the first modal pattern has been identified, we go back to

the stage that analyzed the output of the D-Square factoring. There the pivot profiles are picked up that identify the modal pattern inherent in the second column of the D-Square factor matrix. These pivot profiles are used for computating the second modal pattern. The second modal pattern has the following score values: 5.98, 1.98, 5.98, and 1.98. True values in Figure 1 differ from these again by only .02 raw score points.

Only now after identifying the two modal patterns inherent in the first shape factor, do we return to analyze the second shape factor. D-Squares are computed for profiles loaded on the second shape factor and are then factored. The D-Square matrix is not reported here but the results of factor analyzing the profiles heavily loaded on the second shape-family factor are given in Table 5.

TABLE 5

Rotated Factor Matrix Obtained from Analyzing Transformed Inter-Profile D-Square for Profiles That are Loaded on the Second Shape-Family Factor

Profile	Factor I	Factor II
7	0.0474	17.0355
8	1.8951	16.8841
9	3.6370	16.6268
10	16.6273	3.6378
11	16.8842	1.8951
12	17.0353	0.0470

The third modal pattern is as follows: 17.02, 17.02, 21.02, and 21.02. The fourth modal pattern is: 1.98, 1.98, 5.98, and 5.98. Once again the results differ only .02 score points from the values in Figure 1.

An almost exact solution was obtained for the profiles in Figure 1. But what about clusters of profiles that are much closer in level than in Figure 1, will the procedure still identify them correctly?

Figure 2 presents 12 profiles, similar to those in Figure 1 but the difference in levels between modal patterns is less. The four modal profiles were correctly identified and score values obtained for each profile were:

1.	11.08	7.08	11.08	7.08
2.	5.92	1.92	5.92	1.92
3.	7.08	7.08	11.08	11.08
4.	1.92	1.92	5.92	5.92

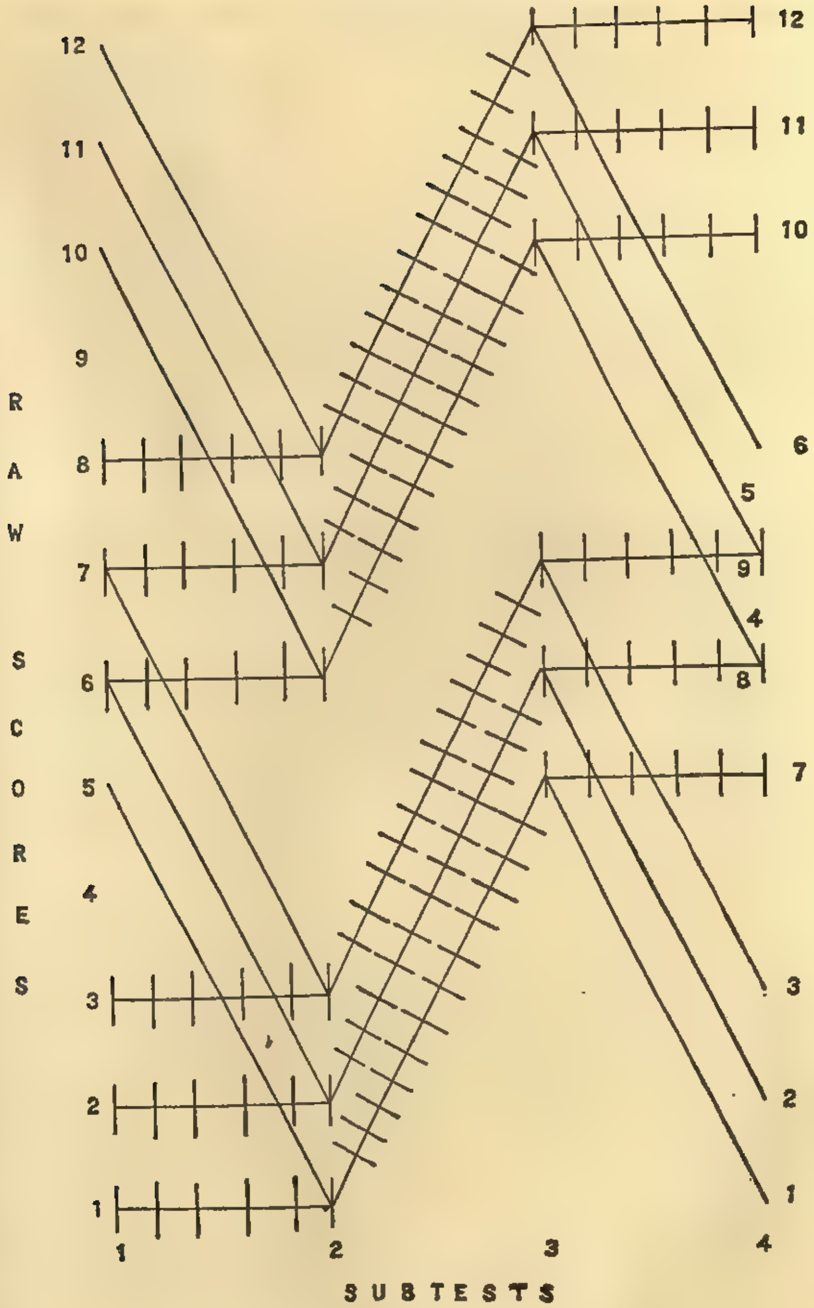

RAW SCORES
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
12
11
10
6
5
9
4
8
7
3
2
1
SUBTESTS

Figure 2.

Figure 3 presents 14 profiles with four patterns distinguishable on the basis of variance, level and shape. These modal profiles were correctly identified and the subtest score values obtained for each score:

1.	1.87	1.87	8.87	8.87
2.	5.13	5.13	6.13	6.13
3.	26.87	19.87	26.87	19.87
4.	24.13	23.13	24.13	23.13

Figure 4 presents 12 profiles with four patterns much like those in the previous figure, except that all have the same average level. These 4 modal profiles were correctly identified and the subtest score values obtained for each were:

1.	1.87	1.87	8.87	8.87
2.	5.13	5.13	6.13	6.13
3.	8.87	1.87	8.87	1.87
4.	6.13	5.13	6.13	5.13

It may never be possible to match all profiles to a few modal patterns, but the percentage of indeterminate cases might be reduced by identifying which tests are most important in the modal patterns. That is, the reciprocal of the average D^2 for each subtest of the profiles in the original cluster can provide weights to be used to establish relationships between undiagnosed *S*s in the sample and the modal patterns. When near zero weights are assigned to some subtests by the program, the comparison is reduced to using the remaining subtests for diagnostic *signs* rather than modal patterns (profiles).

It seems desirable to mention that all modal patterns must be cross-validated before they can be presumed to have diagnostic value, since random sampling error may have contributed greatly to their formation. Also, different patient samples used in cross-validating may disclose additional patterns not sampled earlier. Features irrelevant to diagnosis, such as sex and education, also may cause a systematic separation of modal patterns, and it may prove necessary to propose different modal patterns associated with psychiatric disorders for different classes of people.

No systematic proposal for determining the fit of a *S* to modal patterns is made here. It is supposed that visual inspection and comparison of profile and patterns will suffice for clinical purposes. For more quantitative matching of a profile to patterns the simple statistics of du Mas (1963) would be appropriate.

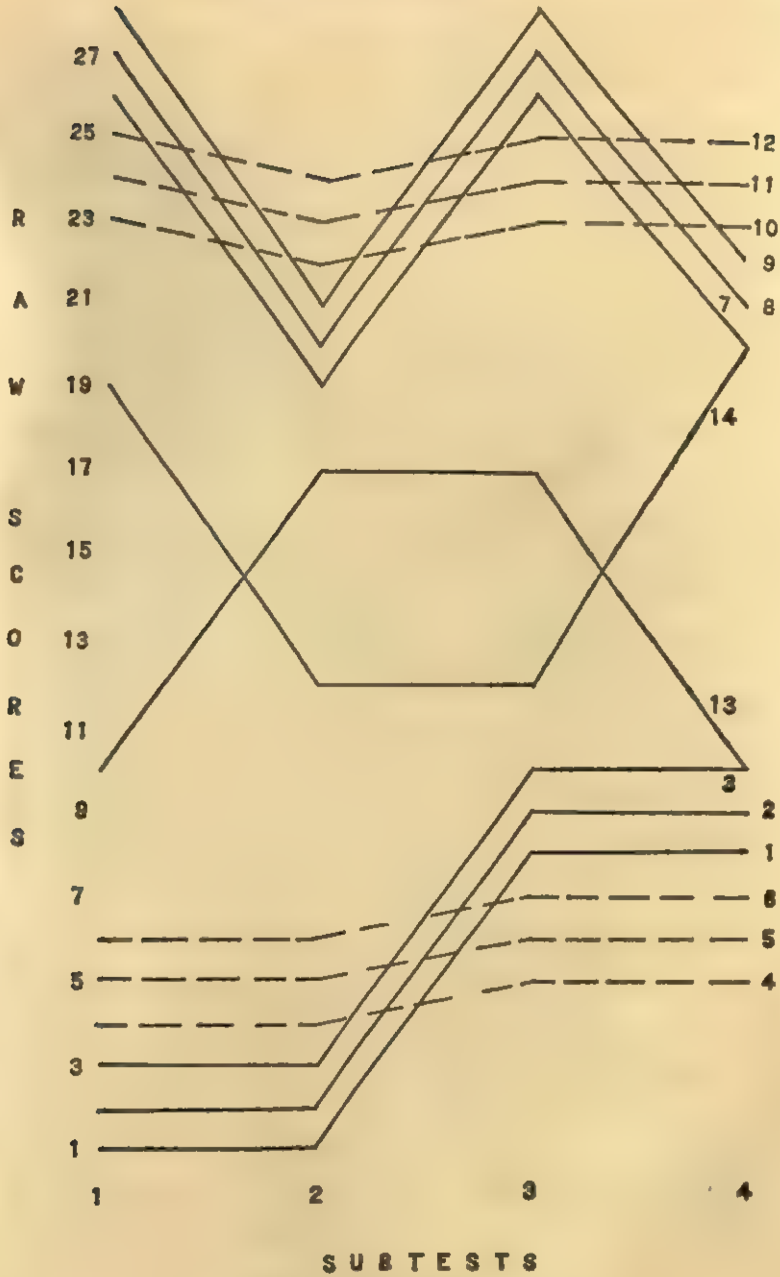

RAW SCORES
27
25
23
21
19
17
15
13
11
9
7
5
3
1
12
11
10
9
8
7
14
13
3
2
1
6
5
4
1 2 3 4
SUBTESTS

Figure 3.

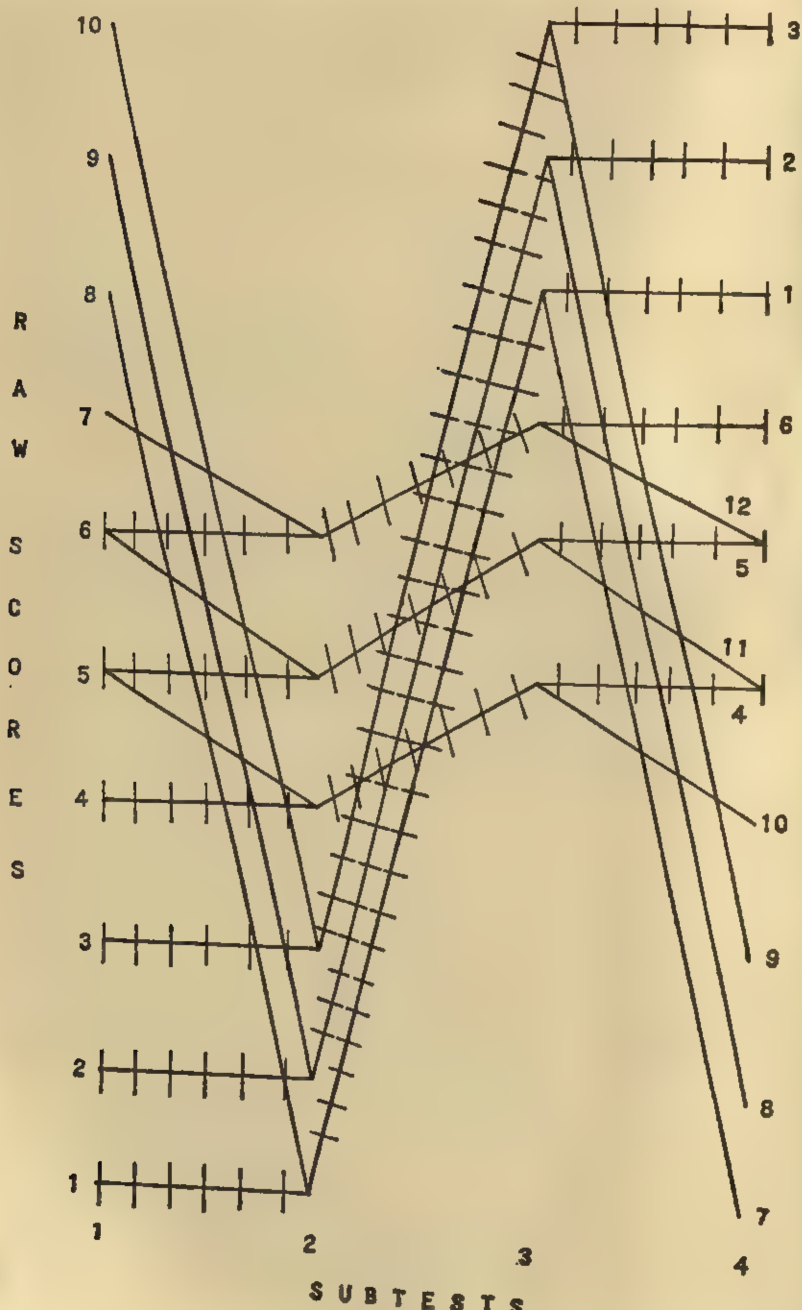


Figure 4.

REFERENCES

Cattell, R. B. "r_p and Other Coefficients of Pattern Similarity." *Psychometrika*, (1949) XIV, 279–298.

Cronbach, L. J. and Gleser, C. G. "Assessing Similarity between Profiles." *Psychological Bulletin*, (1953) L, 456–473.

du Mas, F. M. "The Coefficient of Profile Similarity." *Journal of Clinical Psychology*, (1949) V, 123–131.

du Mas, F. M. "Quick Methods for the Analysis of the Shape Elevation, and Scatter of Profiles." *Journal of Clinical Psychology*, (1953) IX, 345–348.

Haggard, E. A., Chapman, Jean P., Isaacs, K. S., and Dickman, K. W. "Intraclass Correlation vs. Factor Analytic Techniques for Determining Groups of Profiles." *Psychological Bulletin*, (1959) LVI, 48–57.

Mahalanobis, P. C. "On the Generalized Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, (1936) XII, 49–58.

Matthews, C. G., Guertin, W. H., and Reitan, R. M. "Wechsler Bellevue Subtest Rank Orders in Diverse Diagnostic Groups." *Psychological Reports*, (1962) XI, 3–9.

Nunnally, J. "The Analysis of Profile Data." *Psychological Bulletin*, (1962) LIX, 311–319.

Overall, John E. "Note on Multivariate Methods for Profile Analysis." *Psychological Bulletin*, (1964) LXI *61*, 195–198.

Saunders, D. R. and Schueman, H. *Syndrome Analysis: An Efficient Procedure for Isolating Meaningful Subgroups in a Non-Random Sample of a Population.* Paper read at Psychonomic Society, St. Louis, September 1962.

Wechsler, D. *The Measurement of Adult Intelligence.* (3rd ed.) Baltimore: Williams and Wilkins, 1944.

Wechsler, D. *The Measurement and Appraisal of Adult Intelligence.* (4th ed.) Baltimore: Williams and Wilkens, 1958.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

ELECTRONIC COMPUTER PROGRAMS AND ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara

JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

IN view of the tremendous advances that have been made in the adaptation of electronic computers and accounting machines to the processing of statistical data, sections of the Spring and Autumn issues of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT are devoted to the publication of such programs as are appropriate to psychometric procedures. Programs relevant to such problem areas as factor analysis, item analysis, multiple regression procedures, the estimation of the reliability and validity of tests, pattern and profile analysis, the analysis of variance and covariance, discriminant analysis, and test scoring will be considered. Customarily a program should be expected not to exceed six or eight printed pages. Manuscripts of four or fewer printed pages are preferred. Each manuscript will be carefully reviewed as to its suitability and accuracy of content. In some instances an accepted paper may be returned to the author for possible revisions or shortening. The cost to the author will be fifteen dollars per page for regular running text. The extra cost of the composition of tables and formulas will be added to the basic rate. Manuscripts received up to November first will be considered for the Spring issue; manuscripts received between then and May first will be considered for the Autumn issue.

All correspondence should be directed to

William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
University, California

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

MONTE CARLO F-II: A COMPUTER PROGRAM FOR ANALYSIS OF VARIANCE F-TESTS BY MEANS OF PERMUTATION

FRANK B. BAKER
University of Wisconsin
AND
RAYMOND O. COLLIER, JR.
University of Wisconsin

An exact analysis of variance F-test can be obtained through a permutation procedure due to Fisher (1935), and from a theoretical point of view the normal theory F-test is an approximation to this exact test. Although a large number of computer programs exist for the usual normal-theory analysis of variance and its associated F-tests, only a single special purpose program has been reported for the F-test under permutation (Baker and Collier, 1961). The present paper describes a general purpose program for obtaining the empirical distributions of variance ratios under permutation for a wide class of experimental designs. In order to clarify the nature of this computer program, a brief description of the permutation procedure is presented in the paragraph below.

Under the permutation procedure the distribution of a particular variance ratio, say F_t, can be produced from just the data collected in a single experiment. The distribution is obtained by permuting the sample data in accordance with the constraints imposed by the experimental design and by computing F_t for that permutation pattern. The data are then repermuted, F_t recalculated, and the process repeated until a sufficient number of F_t's have been obtained to form the empirical distribution of F_t. One then defines a critical region on this distribution, say the upper five percent of the cases, and if the value of F_t yielded by the sample data falls in this region,

the corresponding null hypothesis is rejected. There are two advantages to using the F-test under permutation. One, only the data at hand are necessary to form an exact test of the null hypothesis. Two, there are no underlying assumptions of normality such as are required by the usual analysis of variance procedures. For additional discussion of the permutation approach, the reader is referred to Scheffé (1959, Chapter 9) and Kempthorne (1952).

The Program

In order to produce an efficient computer program for the above procedure, the program was broken into sections A, B, and C which are under control of a higher level executive routine. Section B permutes the data, performs the analysis of variance, and yields the empirical F's. Section A was assigned a number of tasks which adapt the internal configuration of section B to the particular data and analysis employed. These tasks are performed only once; hence were programmed as a separate section. Section C is a post-processing routine which arranges the obtained F's into frequency distributions and produces the output listings. Generality in the analysis of variance (ANOVA) aspects of the program was obtained by using a general purpose ANOVA program due to Jennrick (1961), which can handle any balanced fully replicated or nested design with up to eight factors involving an equal number of observations per cell of the design. In order to provide complete generality, it was necessary to design a permutation program with generality matching that of the ANOVA program. The permutation scheme consists of two programs, one which translates a set of codes specifying the permutation to be performed (DECODE) and one which actually permutes the data (PERMUTE). The operation codes implemented were as follows:

- * —signifies blocking. The array is separated into blocks, each of which is processed separately.
- \$ —means permute over. The observations to the left of the \$ are permuted over the cells on the right.
- ( )—signifies a set. The letter within the parentheses defines the size of the set. A set is a group of observations to be permuted as a logical entity.
- + —signifies then. Used to separate successive stages of permutation.

• —means end of specifications.

For purposes of illustration, let us define an observation as X_{ijk} and employ a randomized block design with I blocks, J treatments, and K observations per cell. The permutation procedures required by this design are as follows:

1. The JK observations within a given block be permuted.
2. This process be repeated separately for each block in the design.

Such a procedure can be specified using the operation codes by $JK\$JK{*}I.$, which means permute the JK observations into JK cells and repeat this for I blocks. In the case of a design which involves two stages of randomization, let us define an observation as X_{ijkl} with I replications, J blocks, K treatments, and L observations per cell. A permutation specification could be $KL\$KL{*}IJ + (L)\$JK{*}I.$, in which case the KL observations are permuted into KL cells and this is done for IJ blocks. The second stage consists of treating the sets of L observations as a logical unit and of permuting the sets of size L into JK cells for each of I blocks. Given a permutation specification such as one of the above, the subroutine DECODE uses the numerical values corresponding to the levels of the index letters to compute the locations of the corresponding arrays, as well as employs the operation codes to establish the permutation to be performed. The DECODE routine then sets indices, initializes arrays, and prestores the routine PERMUTE which actually does the permuting of the data. Under this scheme the decoding operation is performed only once in section A. Thereafter, the permute routine properly manipulates the data.

The separation of the specification or prestore stage from the computational stage is also maintained in the ANOVA programs. A subroutine named ANOVA-A in section A reads in all the information related to specification of the analysis of variance, the error terms to be used, and the variance ratios to be computed. The information generated from these specifications is used to prestore the program ANOVA-B in section B which computes the variance ratios desired. Thus, once the configurations of the PERMUTE and ANOVA-B routines have been established, they merely permute, compute, and store the F's until the desired number of values have been created.

Output

When the proper number of F's have been computed and stored, section B terminates and the post-processing routines of section C are entered. Section C reads in a set of cards specifying the output desired and then processes the F's provided by section B in accordance with these specifications. The output from section C consists of the following information for each F-distribution specified:

1. A grouped frequency distribution of the empirical F's with the midpoints of the intervals being the 9 decile points and $P_{.95}$, $P_{.975}$, and $P_{.99}$. The normal theory F's corresponding to these points are also given.
2. The cumulative proportion from the upper end of the empirical distribution to the lower end.
3. The average error mean square for each variance ratio.

In addition, the complete ANOVA table for the original sample data is presented as are the original data.

Capabilities and Limitations of the Program

There are a large number of specific "do's" and "don't's" given in the program manual and those presented below are given only to provide the reader with some sense of the scope of the program.

1. Randomization patterns and ANOVA can be specified for a single variable possessing up to eight subscripts.
2. Up to 20 stages of randomization can be specified.
3. The product of the number of levels of the indices must be less than 2500.
4. The number of variance ratios specified cannot exceed 20.
5. The number of sums of squares pooled to form an error term cannot exceed 20.
6. The number of samples generated within a computer run is $\leqq$ 1000.
7. An option exists whereby non-null treatment effects can be added to the permuted data prior to computing the F's. Thus, one can obtain the power of the F-test under permutation for the non-null treatment effects applied.
8. The program written in Control Data FORTRAN 63 contains two short machine language subroutines.
9. Systematic accounts as to running times have not been main-

tained, but for small designs the times are quite short. For example, a randomized block design, three blocks, three treatments, and six observations per cell, requires four minutes for a run of 1000 permutation patterns.

Summary

The permutation model devised by Fisher is yet another example of a powerful statistical technique which has been avoided because of its computational labor. Since the digital computer has removed nearly all restrictions imposed by computational labor, techniques such as the permutation procedure should become an integral part of the statistical tools available to the educational and psychological research worker.

REFERENCES

Baker, F. B., and Collier, R. O. "Analysis of Experimental Designs by Means of Randomization, a Univac 1103 Program," *Behavioral Science,* VI (1961) 369.

Fisher, R. A. The Design of Experiments. Edinburgh: Oliver and Boyd, 1935.

Jennrick, R. I. "1604 Analysis of Variance—04 WISC ANOVA," unpublished manuscript, Computing Center, University of Wisconsin, 1961.

Kempthorne, O. *The Design and Analysis of Experiments.* New York: John Wiley and Sons, 1952.

Scheffé, H. *The Analysis of Variance.* New York: John Wiley and Sons, 1959.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

A SET OF SOCIOMETRIC FORTRAN II ROUTINES[1]

LEOPOLD O. WALDER
University of Maryland

The processing and analysis of peer-rating data represents an ideal use of computers. This work when done by human clerks not only is expensive in time and boredom but also is full of error. The present paper is an updating of a report of a system devised for the IBM 650 (Walder, Greene and Lefkowitz, 1962). This set of routines for the IBM 7094 does more than a system of programs (see also Toigo, 1962) for the 650 did; this is the result of the greater capacity of the 7094. This sociometric package was written (coded)[2] in FORTRAN II for the IBM 7094 with an off-line IBM 1401 used as a card-to-tape converter and as an off-line printer.

A set of 11 subroutines was written to perform various functions. A main program (without the DIMENSION and COMMON statements) which calls all the subroutines is listed here:

```
  CALL RDNAJU
1 CALL ZEROMX
  CALL ZSUMS
  CALL RDJCDS
  CALL SUMMX
  CALL PERCNT
  CALL PRINMX
```

[1] Computer time given by the Computer Science Center of the University of Maryland is gratefully acknowledged here.

[2] A source program listing which includes operating instructions has been deposited as Document number 8625 with the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. A copy may be secured by remitting $1.25 for 35 mm. microfilm or $1.25 for photoprints.

```
CALL CHOICE
CALL CHOSEN
CALL RECIPR
CALL IGNORE
GO TO 1
END
```

SUBROUTINE RDNAJU. This subroutine reads the cards containing the names and 2-digit numbers of the members of the group into arrays used later for the output from subroutines SUMMX, PRINMX, CHOICE, CHOSEN, RECIPR, and IGNORE.

SUBROUTINE ZEROMX. The matrix elements which store the choice (1) or non-choice (0) of the i-th judge concerning the j-th object are cleared to zero. This subroutine could be bypassed if two or more matrices were to be summed.

SUBROUTINE ZSUMS. The row sums of the judge scores and the column sums of the object scores are cleared to zero. This subroutine could be bypassed if these scores were to represent sums from more than one item matrix.

SUBROUTINE RDJCDS. This subroutine reads cards containing the item wording and then judge cards (each containing the number of the judge, the item number, and the numbers of those group members he nominated). A one is added to the matrix element corresponding to the i-th row (number of the judge) and to the j-th column (number of the group member nominated). The item number is monitored to avoid mixing data from several items.

SUBROUTINE SUMMX. Column (object) sums and row (judge) sums are calculated. The row and the column sums are calculated and printed, labelled with the names and numbers of the group members.

SUBROUTINE PERCNT. For each sum a corresponding percentage is calculated. The column percentage = (column sum) / (number of judges × number of items). The row percentage = (row sum)/(number of objects × number of items).

SUBROUTINE PRINMX. The matrix is printed. At the far left, the rows are labelled with the names and numbers of the group members; at the top, the columns are labelled with the numbers of the group members.

SUBROUTINE CHOICE. This prints the names of those objects

chosen by this judge. (This reproduces the response sheet of each judge).

SUBROUTINE CHOSEN. This prints the names of those judges choosing this object.

SUBROUTINE RECIPR. This prints the names of those who reciprocated choices.

SUBROUTINE IGNORE. This prints the names of those who reciprocated non-choices.

This system yields the output needed to detect a classroom's choice pattern. At present the author is developing a response sheet acceptable to the IBM 1232 optical reader which will bypass the manual key-punching of cards.

REFERENCES

Toigo, R. "An IBM 650 Computer Program for the Evaluation of a Reciprocation Index Utilizing Classroom Data Based upon Unlimited Choices Obtained under a Single Sociometric Criterion." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 613–615.

Walder, L. O., Greene, H. E., and Lefkowitz, D. D. "A Method for Deriving 'Flexible' Sociomatrices from Response Forms Appropriate to Children in the Third Grade." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 187–191.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

A FORTRAN ITEM ANALYSIS PROGRAM FOR ITEMS SCORED ON A CATEGORICAL OR INTERVAL BASIS[1]

RICHARD L. GORSUCH
Vanderbilt University

Psychological and achievement tests are usually scored by either of two methods: (1) categorical scoring where one and only one of several possible answers is given credit as "correct," or (2) interval scoring where each of the different possible answers to the item receives varying degrees of credit. An example of the former is the multiple choice achievement test. Of the four or five possible answers the individual may select, one is scored as correct and all others receive no credit. An example of the latter, interval scoring, is the 16 PF (Cattell and Eber, 1964). On this personality test, a person can choose any one of three answers and one extreme answer receives a score of zero, the middle answer a score of 1, and the other extreme answer a score of 2. Another example of categorical scoring is found in the semantic differential (Osgood, Suci, and Tannenbaum, 1957); on each item, the person marks any one of several different positions on a line and receives a score according to how far his "x" is from a given end of the line.

Whereas programs for various computers are available to analyze categorically scored items (Baker, 1963; Dick and Spencer, 1965; Jaspen, 1962; Jones, Pullias, and Michael, 1965; Nichols and Tetzlaff, 1965; Saunders, Gaddis, and Michael, 1962), none appear to be readily available for items scored on an interval basis. Furthermore, few programs make provisions for one or more external criteria be-

---

[1] The development of the program presented in this paper was supported by a grant (HD947) to C. D. Spielberger from the National Institute of Child Health and Human Development, United States Public Health Service.

ing included in the item analysis program even though this is often desirable. Therefore, it was necessary to develop a FORTRAN program for the IBM 7072 which would score and item analyze tests where the items are scored by either a categorical or an interval method. Provisions for evaluating items against external criteria were also included.

Characteristics of the Program

The program was written in basic FORTRAN for the Vanderbilt University IBM 7072 which is fed by an IBM 1401 satellite computer. The IBM 1401 converts the data from cards to magnetic tapes for input and processes both punched and printed output. The program was constructed to run under the Vanderbilt Automatic Monitoring Operating System, but it can be adapted to almost any FORTRAN system. (Source deck listings are available from the author.)

For each item being evaluated, the program calculates the percent of the sample who use each of the possible responses to that item. For items scored on an interval basis, the mean and standard deviation of the responses are provided. Since the mean for categorically scored items is the proportion of the sample answering with the correct response and the standard deviation is a function of that proportion [i.e., $p\ (1\text{-}p)$], both are essentially contained in the table giving the percent responding with the correct response. The space normally used on the page printout for the means and standard deviations is therefore, in the case of categorical items, used to print out the biserial part-remainder correlation of each categorical item with the total score from the remaining items.

In addition to the biserial correlation for the categorical items, the product-moment part-remainder correlation of the item with the total score received on the remaining items is given for both types of items to facilitate internal-consistency evaluation. For item evaluation with external criteria, the product-moment correlation of the item with scores on each external variable is also calculated. The product-moment formula, used for both external and internal evaluation, gives phi, point-biserial, point-polyserial (Jaspen, 1946, 1965), or the usual Pearson product-moment correlation depending on the nature of the data. If both item and criterion are categorical, the result is a phi coefficient. If the item is categorical and the

criterion is continuous, the result is a point-biserial coefficient. The coefficient is of the point-polyserial family when the item has several scoring itervals but the criterion is continuous. All of these coefficients are estimates of the usual Pearson product-moment coefficient without the assumption of a normal distribution of scores of the parent population on the continuum underlying the categories.

The program presents the means and standard deviations of the test scores and of the criteria scores. The internal consistency reliability of the test is calculated by the alpha coefficient (Cronbach, 1951) which is KR-20 in the special case of the categorical items. The program also prints out the number of subjects on which the item analysis is based and a job identification. If desired, each person's raw score will be punched into a card which also contains up to 10 columns of identifying material from his data card.

The input is presently limited to tests having a maximum of 250 items where each item response is read in as a one-digit character (i.e., item scores can only vary from 0 through 9). Up to five criterion variables may also be included for item evaluation. The number of subjects is unrestricted by the program.

Preparation of the Data

A control card follows the program deck with its header cards (which give the Vanderbilt monitor system the user's budget number and other essential information). The control card is punched on an IBM 80 column card as follows:

Columns	*Entry*
1–3	Number of items in the test (up to a maximum of 250)
6	1 for categorical scoring, 2 for interval scoring
9	The minimum score any subject can receive on an item (0 for categorical scoring)
12	The maximum score any subject can receive on an item (1 for categorical scoring)
15	Number of other variables with which each item is to be correlated (up to a maximum of five)
18	1 if punched scores for each subject are *not* desired (left blank if punched scores are desired)
21–80	Job identification (determined by user)

Following the control card is a card or cards containing the scoring key for the particular test being scored. As many cards as necessary are used. The first scoring key card contains the scoring

instructions for items 1 through 80, the second for 81 through 160, and so on. The scoring key cards allow the data to be punched without regard to the direction of scoring of individual items. Beginning in column one, a number is punched for each item in consecutive columns of the card(s). If the scoring is categorical, each column is punched with that category (0 to 9) which indicates the correct answer for that particular item. If interval scoring is used, each column is punched with a 0 unless the scoring is to be reversed. If the scoring for a particular item is to be reversed from the way in which it was punched, a 1 is put in the appropriate column of the scoring key card. For example, on a Likert attitude scale a person may mark any number from 1 to 5 to indicate the degree of his agreement with an attitude statement, which may be positively or negatively related to the attitude being measured. Let us assume that the first column of the scoring key card contains a 0 and the second a 1 to indicate the direction of scoring. Complete agreement with item 1, indicated by a 5 on the person's data card, would then add five points to his total score. Complete agreement with item 2, also indicated by a 5 on the data card would, however, indicate a response against the attitude being measured. The scoring would be reversed and therefore only one point would be added to his total score.

The next card after the scoring key cards contains the FORTRAN format statement describing each person's data. The format begins with a double A field, but all data fields are in floating point notation. For example, the following format might be used for a 100 item test with three criterion variables: (4X, 2A3, 10X, 50F1.0/20X, 50F1.0, 3F5.2). The data cards then follow. Each person's criterion scores are punched after that person's item scores. After one person's data are read, the second A field is checked by the program; if it is blank, the program assumes that the last person's data have already been read. Therefore, the last subject is followed by as many blank cards as are required for one observation to tell the program that the last set of item responses has been read.

Any number of sets of data can be scored on the same pass by repeating the above sequence of control card, scoring key card(s), format card, data cards, and blank end-of-the-data card(s). A blank card after the blank end-of-the-data card(s) indicates to the program that the last set of data has been processed and the run is completed.

REFERENCES

Baker, F. B. "Generalized Item and Test Analysis Program: A Program for the Control Data 1604 Computer." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXIII (1963), 187–190.

Cattell, R. B. and Eber, H. *The 16 Personality Factor Questionnaire* (3rd Ed.). Champaign, Illinois: Institute of Personality and Ability Testing, 1964.

Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, XVI (1951), 297–334.

Dick, W. and Spencer, R. E. "An Application of Computer Programming to Test Analysis and Item Analysis." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXV (1965), 211–215.

Jaspen, N. "Serial Correlation." *Psychometrika*, XI (1946), 23–30.

Jaspen, N. "Self-scoring Item Analysis Procedure for the IBM 1620," EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 595–598.

Jaspen, N. "Polyserial Correlation Programs in FORTRAN." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXV (1965), 229–233.

Jones, R. A., Pullias, C., and Michael, W. B. "An IBM 1401 Computer Program for Item and Test Analysis." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXV (1965), 217–219.

Nichols, R. C. and Tetzlaff, W. "Test Scoring and Item Analysis Programs." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXV (1965), 205–210.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.

Saunders, Sarah G., Gaddis, L. W., and Michael, W. B. "An IBM 650 Program for Item Analysis of Dichotomized Variables." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XXII (1962), 171–176.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

SCORING TEST BATTERY: A PROGRAM FOR THE IBM 7094

ARIEH LEWY and WILLIAM CRAWFORD
University of Illinois College of Medicine

This program is designed to score a complex test battery and to provide individual scores, test statistics, and item data for each subtest separately and for the battery as a whole. It may also be used for scoring a test which is not divided into subtests. Both printed and punched output are produced.

A great saving in computer-time is accomplished by scoring all items in one pass through the computer after which item scores are assigned to subtest categories. Therefore it is not necessary to search core storage repeatedly or to rewind tapes to obtain scores for each subtest.

I. Restrictions

The program is designed to score a test composed of a maximum of 500 items and 26 subtests. No item may be assigned to more than three subtests. The number of alternatives may range from two to nine and need not be the same for every item. There may be no more than 20 item response cards per subject (S).

II. Options

Subtest classification

Each item may be assigned to as many as three subtests. This feature permits cross-classification of items so that any item may contribute to more than one subtest score. A separate score is computed for each subtest, and the total score is accumulated by count-

ing each item only once regardless of the number of subtests to which is contributes.

Weighting

Each item may be assigned a different weight, but the weight for an item must be the same in every subtest to which it contributes. The subtest scores can also be weighted and combined into a composite weighted total score. Both options can be used simultaneously to obtain subtest scores based on weighted items and a total score based on weighted scores and weighted items.

T Scores

In addition to the raw and percent scores produced as standard output, *T*-Scores may be obtained for each of the subtests and for the total test.

Card Sequence Control

This option may be used to compare identification numbers only, or to check both identification numbers and card sequence for each *S*. The sequence control statements compare identification numbers as well as the card sequence numbers for each *S* to determine whether the proper number of cards is present and whether the cards are in correct order. This option is especially valuable if examinations are to be scored for purposes of grading or of individual selection, since incorrect card sequencing or missing data could radically alter a student's score and since it is possible that such an error could go undetected if an automatic check were not available. If an error is found, a message (CHECK CARD SEQUENCE) is printed beside the scores for that *S* and scoring continues without interruption.

The card numbers must be positive integers in increasing order for each *S* separately. However, the integers need not be consecutive nor need they be the same for each *S*. Therefore sequence control can be used when only certain cards for each *S* are selected for special processing.

III. Input and Output

Input is in the form of punched cards which contain an *S* identification number, card sequence control numbers, and raw item re-

sponses. There may be as many as 20 cards per S, and each card may contain up to 73 item responses, provided that the total number of responses per S does not exceed 500. One variable format statement governs the reading of all data cards. This statement must be prepared so that it continues over as many cards as there are data cards for each S.

Printed Output:

(1) Individual Scores
   (a) raw scores for the total and each subtest (standard),
   (b) percent scores for the total and each subtest (standard),
   (c) weighted scores for the total and each subtest (optional),
   (d) T-scores for the total and each subtest (optional);
(2) A frequency distribution of scores on the total test and on each of the subtests;
(3) A complete intercorrelation matrix for all scoring categories;
(4) The proportion of Ss selecting each alternative for every item;
(5) Item statistics computed with respect to the total and every subtest:
   (a) difficult,
   (b) item standard deviation,
   (c) point biserial correlation,
   (d) index of reliability; and
(6) The following test statistics for each subtest and the total:
   (a) maximum possible score,
   (b) mean and its standard error,
   (c) standard deviation and its standard error,
   (d) skewness and its standard error,
   (e) kurtosis and its standard error,
   (f) reliability (KR-20)

The intercorrelation matrix and all subtest and total scores may be obtained as punched output. Punching is controlled by a variable format statement so that the output may be obtained in a format which can be used immediately for research or administrative purposes.

IV. Computing Procedures

Items are assigned to subtests by a vector of two-digit fixed point numbers which indicates the subtest number of items appearing on corresponding columns of the data cards.

Item weights are assigned by a vector of two-digit floating point numbers. When items are differentially weighted both the total score and the subtest scores are influenced by the weighting factors for each item.

V. Computing Time

All items are scored in one pass through the computer before item scores are assigned to the total and to subtest categories. The program is very economical, since it is not necessary to waste valuable computer time by repeatedly searching core storage or rewinding tapes to obtain scores for each subtest. For example, a test consisting of 361 items classified into 15 subtests taken by 196 students was scored in 2.96 minutes.

Summary

A rapid and efficient program for the IBM 7094 has been developed which scores and provides complete item and test data for a complex battery of multiple choice tests. It also provides options for card sequence control, item weighting, and score weighting. Input is in the form of punched cards containing student identification numbers, card sequence control digits, and raw item responses. Output consists of punched and printed scores and of printed item and test data.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 1, 1966

A 7094 FORTRAN PROGRAM FOR THE COMPUTATION OF TETRACHORIC CORRELATIONS[1]

STEVEN G. GOLDSTEIN, JAMES D. LINDEN
AND DAVID A. STUDEBAKER
Purdue University

A variety of computer programs in use today depends upon correlation matrices for the solution of problems. Many of the original data utilized in the construction of these matrices are in dichotomous format. Because of the unavailability of a program for the computation of the tetrachoric correlation, Pearson product-moment correlations generally have been employed which provide spurious estimates of relationships among the data involved. As a consequence of the need for a program that would compute tetrachoric correlations to provide more appropriate item reliability information and appropriate input for factor analytic work, the authors developed the program reported here.

Characteristics of the Program

Preceding a listing of the tetrachoric correlation matrix, the program listing provides each subject's total score and a frequency distribution of such scores along with the mean, median, variance, and standard deviation of the distribution. The analysis utilizes a two-by-two contingency table peculiar to each set of dichotomous variates represented in the resultant matrix. All possible off diagonal variate combinations are considered. In each case, an *ad/bc* ratio is calculated as described by Downie and Heath (1959). This

[1] The development of this program was supported in part by Contract No. N62269-2670, Naval Air Development Center, Johnsville, Pa., Principal Investigator, Donald R. Brown.

resultant ratio is compared to the estimates of values for the tetrachoric correlation statistic determined by Davidoff and Goheen (1953), and each value selected is either listed or both listed and punched in a variable format. One option arbitrarily provides unity values on the main diagonal of the matrix. Another option substitutes a high-low median split of the total subject scores for one dimension of the contingency table and provides main diagonal tetrachoric estimates of the correlation of a given variate with the total score.

Because the scanning and storing operations demanded by the program may exceed the storage capacity of ths system, the program provides for the temporary storage of information on tape until the monitor calls forth the appropriate information for the output tape.

This program is written to be run under the IBSYS monitor.

Job Deck Set-Up

First Card: This is generally the system identification card and is peculiar to each individual user's system. The program deck immediately follows this card.

Second Card: (Title Card) Any title desired may be punched into columns– 1–72 and is printed out at the top of each listing page.

Third Card: (Problem Card)

Columns 1–3. The number of subjects (less than 171).

Columns 4–6. The number of items or dichotomous variates (less than 121).

Column 7. A "0" punch arbitrarily yields unity values along the main diagonal of the matrix. A "1" punch substitutes the correlation of each item with the total score for main diagonal unity values.

Column 8. A "0" punch will list the contingency ratio resultants. A "1" punch ignores such listing.

Column 9. Punched BCD output is obtained with a "0" punch, whereas a "1" punch provides only the tetrachoric correlation listings.

Column 10. The number of data input format cards.

Columns 11, 12. The number of correlations per BCD card (less than 12).

Fourth Card: (Data Input Format Cards) The program demands an identification number for each subject on the first card of each set of subject cards. Thus the format card must begin with an integer identification number. Subsequently, any number of columns may be omitted and data read as desired. The entire card should be in Integer format.

Data Cards: Data cards follow immediately. They are set up in such a way that each subject set is complete before the next subject set begins. Data must be punched either "1" or "0." They may extend for the entire card field.

Punched Output Cards

These cards are identified by the matrix row and then by the ordinal position of the card in the make-up of the row. The output which is in Floating Point format (F5.2) is fixed by the program. The punches in columns 11 and 12 of the problem card indicate the number of correlations contained on the card. The deck of such BCD cards may then be used as an input matrix for a factor analysis or for any program that requires correlation matrix input.

REFERENCES

Davidoff, M. D. and Goheen, H. W. "A Table for the Rapid Determination of the Tetrachoric Correlation." *Psychometrika*, XVIII (1953), 115–121.

Downie, N. M. and Heath, R. W. *Basic Statisitcal Methods*. New York: Harper & Brothers, 1959.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

A PROGRAM FOR COMPUTING CANONICAL CORRELATIONS ON IBM 1620[1]

E. ROSKAM
Eindhoven Institute of Technology, Netherlands

Introduction

Canonical correlations are called for (e.g.) when relationships are investigated between a set of predictor variables and a criterion consisting of a number of variables that cannot on a priori grounds be combined into a single measure. Partial regression weights can be found which maximize the correlation between the weighted sum of predictors and the weighted sum of criteria. Besides these, different sets of weights can be found providing independent weighted sums, thus giving 2nd, 3d etc. canonical correlations. The maximum number of independent sets of weights is equal to the number of predictors or to the number of criteria, whichever is the smaller. Each set of weights can present valuable information concerning the construct validity of each of the variables separately. It is therefore worthwhile not only to compute the maximum canonical correlation, but also to proceed to the smaller ones, as long as they reach statistical significance.

Two FORTRAN programs for computing canonical correlations have been described by Cooley and Lohnes (1962). One of these computes the maximum canonical correlation only and it does not assess its statistical significance. The other program computes all canonical correlations and also computes chi-square tests for each of them.

---

[1] The program has been tried out successfully on the IBM-1620 in the Mathematical Department of the Eindhoven Institute of Technology. I am indebted to Dr. A. J. Geurts for his kind cooperation and assistance.

For many practical purposes the latter program performs too much, because usually only the first few canonical correlations will be statistically significant. Moreover, its length is prohibitive for use on a small computer.

This paper describes an alternative program, which computes canonical correlations in decreasing order of magnitude and provides statistical tests for each. The number of canonical correlations to be computed is given as input, but it might also be decided by the machine on the basis of a given level of significance.

Rationale

Input to the computer is a super-matrix of correlation coefficients between m_p predictor variables and m_c criterion variables.

$\mathbf{R}_{pp}$ intercorrelations among predictors	$\mathbf{R}_{pc}$ intercorrelations between predictors and criteria
$\mathbf{R}_{cp}$ transpose of R_{pc}	$\mathbf{R}_{cc}$ intercorrelations among criteria

Since it is symmetric, the lower triangular half can be omitted. In the canonical correlations model it is irrelevant which set of variables is considered as predictors. The following mathematical expressions are valid only if the smaller set is considered as criteria.

The i-th canonical correlation (r_i) is the square root of the i-th latent root of the matrix

$$\mathbf{X} = \mathbf{R}_{cc}^{-1}\mathbf{R}_{cp}\mathbf{R}_{pp}^{-1}\mathbf{R}_{pc} \tag{1}$$

The weights ($\mathbf{w}_{ci}$) associated with the criteria are the elements of the corresponding latent column vector normalized to unit length. The weights for the predictors are given by the vector

$$\mathbf{w}_{pi} = \mathbf{R}_{pp}^{-1}\mathbf{R}_{pc}\mathbf{w}_{ci}/r_i, \tag{2}$$

normalized to unit length.

The statistical test of significance for the i-th canonical correlation is given by

$$T_i = -[N - \tfrac{1}{2}(m_p + m_c + 1)] \sum_{j=i}^{j=m_c} \log_e (1 - r_j^2) \qquad \begin{matrix} r_{j+1} < r_j \\ m_c < m_p \end{matrix} \tag{3}$$

where N is the sample size. This is distributed as chi-square for $(m_p - i + 1)$ $(m_c - i + 1)$ degrees of freedom. For details and math-

ematical proofs of (1), (2) and (3), see Cooley and Lohnes (1962), Anderson (1958), and Bartlett (1941).

The computation of the chi-square test seems to involve the computation of all canonical correlations. However, a method which greatly simplifies this labor has been found. For $i=1$, the sum in (3) is equal to the natural logarithm of the determinant[2] of $\mathbf{I} - \mathbf{X}$. For $i=2$ it suffices to subtract $log_e(1-r_1^2)$ from this sum to calculate the chi-square test for r_2, etc. This method permits of testing the significance of each canonical correlation as it is computed in decreasing order of magnitude, without computing all of them separately.

Two requirements are essential to use this advantageous possibility: (a) The availability of a method for assessing the determinant of a matrix without unduly lengthening the computational program; in fact a subprogram for matrix inversion (which is required anyhow) was used which provides the determinant of a matrix as a byproduct.[3] (b) The use of an extraction method which provides the latent roots of (1) successively in decreasing order of magnitude.

One method of computing the largest latent root and vector, which is adequate in most cases, is the standard iterative method (also known as the Hotelling method). Since matrix (1) is *not* symmetric, it is not possible to compute the next largest latent root and vector by means of the extraction or exhaustion method in the usual way[4].

Anderson (1958), however, describes a relatively simple procedure which permits of the consecutive application of the iterative method to subsequent residual matrices. This method consists of first normalizing the criterion weights in such a way that

$$\mathbf{w}_{ci}^{*\prime}\mathbf{R}_{cc}\mathbf{w}_{ci}^{*} = 1 \tag{4}$$

where * denotes the normalized vector. Let now $\mathbf{W}_c^*$ stand for the matrix of normalized criterion weights associated with all canonical correlations.

---

[2] I am indebted to Prof. G. W. Veltkamp for pointing out this relationship. *Proof:* Let λ be a latent root of $\mathbf{X}$. Then $|\mathbf{X}-\lambda\mathbf{I}|=0$; $|-\mathbf{X}+\lambda\mathbf{I}|=0$; $|\mathbf{I}-\mathbf{X}-\mathbf{I}+\lambda\mathbf{I}|=0$; $|(\mathbf{I}-\mathbf{X})-(1-\lambda)\mathbf{I}|=0$; hence $1-\lambda$ is a latent root of $\mathbf{I}-\mathbf{X}$. Therefore $|\mathbf{I}-\mathbf{X}|=\Pi(1-\lambda)$, Q.E.D.

[3] This subprogram is part of the standard software of the Mathematical Department of the Eindhoven Institute of Technology, where it has been developed and tested. Particulars of this subprogram are given below.

[4] Unless by also computing the latent row vector, which is unduly laborious.

As

$$\mathbf{w}_{ci}'\mathbf{R}_{cc}\mathbf{w}_{cj} = 0 \qquad i \neq j$$

therefor:

$$\mathbf{W}_c^{*\prime}\mathbf{R}_{cc}\mathbf{W}_c^{*} = \mathbf{I}$$

or

$$\mathbf{W}_c^{*\prime}\mathbf{R}_{cc} = \mathbf{W}_c^{*-1} \tag{5}$$

Hence, for the i-th canonical correlation

$$\mathbf{w}_{ci}^{*\prime}\mathbf{R}_{cc} = \bar{\mathbf{w}}_{ci}^{*\prime} \tag{6}$$

In this expression $\bar{\mathbf{w}}_{ci}^{*\prime}$ is the i-th row in $\mathbf{W}_c^{*-1}$. Anderson shows that the largest latent root of

$$\mathbf{X}_{i+1} = \mathbf{X}_i - r_i^2\mathbf{w}_{ci}^{*}\bar{\mathbf{w}}_{ci}^{*\prime}, \qquad \mathbf{X}_1 = \mathbf{X} \tag{7}$$

is equal to the second largest latent root of $\mathbf{X}_i$. Hence the (i+1)th canonical correlation can be found by applying the iterative method to $\mathbf{X}_{i+1}$.

An obvious advantage of the iterative method is that it requires a relatively short program, and, secondly, that the computation of canonical correlations can be stopped after a predetermined number or if the next r were below a certain level of statistical significance. Other methods of computing latent roots and vectors of nonsymmetric matrices would then be unnecessarily laborious. The principal disadvantage of the iterative method is that it becomes less reliable and more time consuming in computing the smaller and more nearly equal latent roots. In most cases, however, only the first few latent roots, which usually are well separated, will provide significant canonical correlations, and therefor this disadvantage is not very serious.

Computational Flow

The above procedure has been programmed in FORTRAN for IBM 1620 (memory storage 6000 words). The program is adapted to the requirements of the available FORTRAN system. For instance no "CALL *subroutine*" statements are included. The capacity is 25 predictors and 25 criteria (other combinations are possible).

The computational flow, omitting details, is as follows:

The subprogram *Inverse* operates on a matrix **M** stored in A. This subprogram is composed of two parts. *Inverse part one* replaces the matrix in A by an uppertriangular and a lowertriangular matrix, so that **LU**=**M**. The diagonal elements of **U** are unity and they are not stored.

Next *Inverse part two* replaces the two triangular matrices by the inverse of **M**.

Inverse part one is used for the computation of the determinant of **M**. As the determinant of **U** is unity, the determinant of **L** will be equal to the determinant of **M**. Since **L** is triangular, its determinant is equal to the product of its diagonal elements.

The main program consists of the following steps:

1. $\mathbf{R}_{pp}$ and $\mathbf{R}_{pc}$ are read into the two-dimensional arrays A en C respectively;
2. $\mathbf{R}_{pp}$ (located in A) is subjected to the subprogram *Inverse.* A now contains $\mathbf{R}_{pp}^{-1}$.
3. The product $\mathbf{R}_{pp}^{-1}\mathbf{R}_{pc}$ is computed and stored in the two-dimensional array P.
4. $\mathbf{R}_{cc}$ is read into both A and the two-dimensional array D. A is subjected to the subprogram *Inverse* so that is replaced by $\mathbf{R}_{cc}^{-1}$.
5. The product $\mathbf{X} = \mathbf{R}_{cc}^{-1}\mathbf{R}_{cp}\ (\mathbf{R}_{pp}^{-1}\mathbf{R}_{pc})$ is computed and stored in A. It will be noted that $\mathbf{R}_{pp}^{-1}\mathbf{R}_{pc}$ is preserved in P for the computation of the predictor weights; $\mathbf{R}_{cc}$ is preserved in D for the computation of the residual matrix $\mathbf{X}_{i+1}$.
6. **X** is copied into C. **I**—**X** is placed into A and subjected to *Inverse part one,* and $log_e|\mathbf{I}-\mathbf{X}|$ is computed; this value is referred to as PI.
7. The largest latent root and vector of the matrix stored in C (consecutively $\mathbf{X}_1$,$\mathbf{X}_2$, etc.) is computed by means of the iterative method. Canonical *r,* criterion and predictor weights and chi-square are computed and printed. The program either comes to an end if the required number of canonical correlations is reached, or it transfers to the next step:
8. PI is decreased by $log_e(1-r^2)$ for the chi-square test of subsequent canonical correlations.
9. The criterion-weights are further normalized by dividing them by the scalar $\sqrt{\mathbf{w}_c'\mathbf{R}_{cc}\mathbf{w}_c}$. The normalized vector is multiplied by $\mathbf{R}_{cc}$ to provide $\overline{\mathbf{w}}_c^*$.

10. The residual matrix $\mathbf{X}_{i+1} = \mathbf{X}_i - r_i^2 \mathbf{w}_{c_i}^* \overline{\mathbf{w}}_{c_i}^{*\prime}$ is computed and stored in C. The program transfers back to step No. 7.

The program provides for an automatic STOP in the rare cases that a canonical correlation should equal unity or if one of the inverses does not exist. A STOP is also provided in case the computation of the latent root and vector should not converge.

A complete FORTRAN listing can be obtained from the author, or from the Director of the Testing Bureau, University of Southern California, Los Angeles, California.

REFERENCES

Anderson T. W. *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley, 1958.

Bartlett, M. S. "The Statistical Significance of Canonical Correlations," *Biometrika* XXXII (1941), 29–38.

Cooley, W. W. and Lohnes, P. R. *Multivariate Procedures for the Behavioral Sciences.* New York: John Wiley, 1962.

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

TO ERR IS INHUMAN: EFFECTS OF A COMPUTER CHARACTERISTIC

KENNETH I. HOWARD and ROBERT W. LISSITZ[1]
The Institute for Juvenile Research

High speed computer equipment is in very wide use today, and is playing a particularly important role in the behavioral sciences. With the introduction of low rental "desk-size" computers, smaller research establishments are able to support computer installations. However, computer experts are not always available to these installations for general consultation, and the individual researcher frequently finds it more convenient to write and operate his own programs. This tendency has become more prominent with the proliferation of simplified programing languages.

Truncation Error

Reliance on computers, often to the neglect of basic research techniques (e.g., random groups, reliable rating scales), has been criticized frequently (for example, see the "comment" in the *American Psychologist,* 1964, by Crumbaugh). Most of these criticisms, which have focused on the quality of data collection and preparation for calculations, emphasize that confidence in the results of a project should not be a function of the complexity of the data processing equipment employed. One aspect of machine use which is not generally mentioned in these criticisms, however, is that of the potential effects of "truncation error."[2] Truncation may cause gross

[1] Now at Syracuse University.

[2] Truncation is the loss of digits imposed by the computer to keep a constant size mantissa. For example, if .15992 is multiplied by .12613, the result is .0201707096. When this problem is handled inside the computer (as distinct from a desk calculator) the result would be truncated to .20170709 and an

calculation errors. It derives from a characteristic of the computer system, not from characteristics of the research.

Neely (1965) has done the basic work on the internal documentation of computer programs at the University of Chicago Biological Sciences Computation Center. He has demonstrated marked effects of truncation error and is presently completing a full-length presentation for computer users and programmers. The following comments and illustrations, which bring one aspect of Neely's conclusions to the attention of behavioral scientists, are provided as a caution, particularly for the relatively unsophisticated computer user.

Most fixed word length computer programming systems (the most common type) use a fixed mantissa of eight digits and an exponent (with a base 10) from −99 to +99. Because of truncation the number represented inside the machine as the result of a mathematical calculation is always less than or equal to the actual number. Therefore, whenever a datum is greater than eight digits at input time, or whenever the result of a calculation is a number with more than eight digits (and the integers past the eighth place are not zeros), an error has occurred and these errors are cumulative throughout a problem.[3]

It is evident that truncation error is a function of the number of significant digits, but there is an important implication which is not so obvious—*different formulae (algebraicly equivalent) for the same statistic* (using different steps to arrive at the same end) *will give different values for the same data set.*

Demonstration of Effects of Error

In order to demonstrate this point, a program was written which would calculate the mean, variance, and product moment correlation, each by two formulae. These formulae are shown in Tables 1

adjustment in the exponent would also take place. First, the computer has the number $.15992000 \times 10^{\circ}$ and the number $.12613000 \times 10^{\circ}$, and then performs the multiplication by referring to the proper place in the table and the result is $.20170709 \times 10^{-1}$.

[3] Some computer languages allow the programmer to vary the length of the mantissa; truncation will, of course, still occur, but perhaps with negligible effects (in terms of the purposes of the programmer). If he is using an "assembly" language (e.g. SPS), the programmer not only can utilize more significant digits, but also can obtain an indication of the size of the truncation error by employing a "noise digit" (see Leeson and Dimitry, 1962, p. 167).

and 2. Data were constructed which would most clearly illustrate the effects of truncation error. Seven problems were run for each formula. Each problem had an N of 100, but a different number of significant digits. In each of the seven sets, the numbers were in ascending order in terms of the last three digits.

TABLE 1

Mean and Variance Using Two Formulae for Each

Range of Data Set	Mean[a] $\frac{\Sigma X}{N}$	Mean[a] $\frac{\Sigma X}{N} + \Sigma\frac{X - \bar{X}}{N}$	Variance[b] $\frac{\Sigma(X - \bar{X})^2}{N}$	Variance[b] $\frac{N\Sigma X^2 - (\Sigma X)^2}{N^2}$
I. 1–100	50.5	50.5	833.25	833.25
II. 901–1000	950.5	950.5	833.25	833.25
III. 9001–9100	9050.5	9050.5	833.25	796.00
IV. 90001–90100	90050.5	90050.5	833.25	−2400.00
V. 900001–900100	900050.5	900050.5	833.25	−350000.00
VI. 9000001–9000100	9000046.4	9000050.5	850.06	30000000.00
VII. 90000001–90000100	90000001.0	90000050.0	3283.50	4500000000.00

[a] The correct result ends in 50.5, and the $\bar{X}$ in the second formula was calculated in an initial pass through using the formula $\bar{X} = \Sigma X/N$.

[b] The correct result is 833.25.

As can be seen in Table 1, no difference between the formulae for the mean occurs until the number of significant digits exceeds six. After this point the standard calculator formula yields discrepancies, whereas the other formula continues to give the correct result to eight significant digits. The error occurs much earlier with the calculation of the variance. The "definitional" formula is essentially correct up to, but not including the eight digit data. The standard calculator formula yields incorrect results after the data exceed three significant digits!

The correlation matrices are presented in Table 2 with the results of the *"definitional" formula in the lower half of the matrix* and the results of the *computational formula in the upper half*. Both formulae give the correct answer when data Set I is correlated with data Set II. After that point, the computational formula yields inaccurate results. The "definitional" formula is essentially accurate for data up to, and including, seven digits. All entries in Table 2 *should be* 1.0000, since the last three digits in each data set are ordered from 001 to 100. Entries which deviate from 1.0000 do so because of the effects of truncation error. Table 2 demonstrates, for example, that, using a *computer program* for the ordinary "desk calculator" form-

TABLE 2

Produce Moment Correlations[a] *Using Two Formulae*

Data Set	I	II	III	IV	V	VI	VII
I	......	1.0000	1.0231	.5876	.0475	.0049	.0001
II	1.0000	......	1.0183	.5629	.0205	.0013	.0005
III	1.0000	1.0000	......	.2966	−.2097	−.0518	−.0000
IV	1.0000	1.0000	1.0000	......	−1.4146	−.2609	−.0000
V	1.0000	1.0000	1.0000	1.0000	......	−.0309	−.0504
VI	.9901	.9901	.9901	.9901	.9901	.....	.9798[b]
VII	.5038	.5038	.5038	.5038	.5038	.6202	.....

Definitional Formula (Entries below diagonal)

$$\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

Computational Formula (Entries above diagonal)

$$\frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2-(\Sigma X)^2}\sqrt{N\Sigma Y^2-(\Sigma Y)^2}}$$

[a] The correct result is 1.0000. The $\bar{X}$ in the definitional formula was calculated in an initial pass through using the formula $\bar{X} = \Sigma X/N$.

[b] This was a fortuitous accuracy, which resulted from a conjunction of errors.

ula, the correlation between two identically ordered sets of data one of which consists of three digits (001–100) and the other of five digits (90001–90100) is .5876.

Cautions for Computer Use

The results clearly demonstrate that truncation effects can be responsible for very large errors in statistical calculations.

There are several ways to minimize the effects reported here. It must be kept in mind that a computer system is not just a very fast, efficient, and obedient research assistant. Computers have built-in sets of rules which must be understood and taken into consideration when processing an analysis. Once the decision has been made to utilize a computer, the data should be examined, not only to predict the outcome of the calculations (this is a good partial check on the program), but also to look for suceptibility to truncation error. Data should be limited to the number of significant digits indicated by the accuracy of the measurement operations employed in the research. When one is dealing with data of more than three digits and large N, extra care must be taken, particularly when the data are composed of a constant plus variants. This type of data may be decreased in length by first subtracting the value of the mean (any constant in the data range will be effective) and then proceeding with the calculations. (After the result is obtained the value subtracted could be reinstated.) A variation of this technique was re-

sponsible for the accuracy of our second formula for calculating a mean. It first computed an approximation of the mean, then found the average deviation from this value, finally added this to the approximation, and thus obtained an accurate solution (even for "maximal error" data of eight digits).

A second approach is to choose statistical formulae very carefully. In addition to not being a research assistant, a computer is not just a very fast desk calculator. Formulae most effective for the desk calculator are not necessarily the best for the computer. In general, formulae should be used which keep the value of a calculation, at any one time, as low as possible. For example, formulae should be chosen which minimize the use of exponentiation (e.g., squaring).

In conclusion then, truncation error can play an important part in calculations. However, if the numbers are small and if the researcher is careful, the effects can be minimized. It is clear that computers should not be treated as infallible "black boxes," but should be used judiciously, with informed checks of both input and output.

REFERENCES

Crumbaugh, J. C. "Mathematical Manslaughter by Electronic Computers." *American Psychologist,* XIX (1964), 775–776.

Lesson, D. N. and Dimitry, D. L. *Basic Programming Concepts and the IBM 1620 Computer.* New York: Holt, Rinehart and Winston, Inc., 1962.

Neely, P. M. *Comparison of Several Algorithms for Computation of Means, Standard Deviations and Correlation Coefficients,* Mimeo, Biological Sciences Computation Center, University of Chicago, 1965.

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

Educational and Psychological Measurement
Vol. XXVI, No. 1, 1966

Fundamental Statistics in Psychology and Education (4th Edition) by J. P. Guilford. New York: McGraw-Hill, 1965. Pp. 605. $8.50.

It is always enlightening to see reflected in a new edition of a standard text the effect of a decade's progress on the author's interpretation of his field. To Guilford, the major innovations of the last ten years are obviously an increased emphasis on the probability and distribution function bases of statistics and the development of inferential techniques of markedly greater power, especially for small samples. This view has led to several changes in the content and organization of the fourth edition of Guilford's widely used text.

Most directly, a new chapter devoted to probability and some fundamental ideas about distributions precedes the material on sampling statistics. The old Chapter 9 has been subdivided into two new chapters. The first deals in general terms with estimation and inference and the second with significance of differences. In the latter, increased and well-deserved emphasis is placed on small-sample methods. The treatment of nonparametrics has been expanded to include the Kolmogorov-Smirnov tests, and its emphasis has been changed by treating chi-square in a separate chapter. The chapter on hypothesis testing gives a new treatment of Type II errors and power functions and devotes more space to sample size requirements. For more purely psychological reasons, the chapter on prediction of attributes has been eliminated and its essentials coupled with the accuracy of the prediction chapter. Throughout the book, more mathematical demonstrations and proofs have been placed at strategic points. A number of these are good, but in general they are probably above the heads of mathematically naive students and a bit trivial to the sophisticated ones. In the same vein, the repeated small inaccuracies in stating statistical concepts will prove irritating to more rigorously inclined teachers even though the practice makes some topics more comprehensible to the students.

Guilford has trimmed some excess wood from his first six chapters and improved them in the process. The detailed descriptions of such fundamentals as setting up frequency distributions remain the best source to which introductory students can be sent. The cataloging of facts about measures of central tendency and dispersion

is both highly readable and intelligible. It does seem superfluous in this day when desk calculators are available to everyone and computers to many, to teach the trial balance method of calculating the mean, standard deviation, and correlation coefficient. It is a common view among teachers that the details of these procedures needlessly confuse the main issues. A short discussion of preparing data for simple computer programs to perform these tasks would be much more appropriate.

As one would expect, the chapters on correlation and prediction are excellent, as is the treatment of reliability. Guilford's views on validity are, of course, based largely on factor theory, and while the concept is thoroughly covered from this angle, a number of other ideas tend to be slighted or oversimplified in the process. The final chapter on comparability of scores and norms is a sparkling clear exposition of a topic that is especially prone to opaque presentation.

Viewed from the perspective of the total book, one wonders why Guilford bothers to include a section on analysis of variance. It is clearly a topic of little interest to him and one that is comprehensively covered in a number of standard sources. This chapter, which is uncharacteristically muddy, is marred by a discussion of multiple comparisons that includes nothing more recent than Tukey's gap test. This section should be avoided.

There are a number of other points about the fourth edition that deserve mention. Throughout the text, informative references and suitable outside readings on various topics are conveniently footnoted. The table of the normal curve is much improved. The problem sets are appropriate and useful. The entire book is set in unusually clear and readable type and format, and, however irrelevant to the book as a text, the binding is most colorful and attractive.

Although primarily intended as a text, Guilford's book will certainly serve as a reference volume for many correlation and prediction problems and for its excellent abacs of the phi coefficients and tetrachoric *r*, among others. As a text the fourth edition has one main drawback. To make a perhaps excusable pun, it is a power text. The difficulty level mounts steadily and fairly rapidly throughout much of the book. A student who hesitates is likely to be lost, and a teacher who forges steadily ahead is likely to find himself alone at the end of the course. For the teacher who takes the pains to keep his students with him and still move forward, the fourth edition will provide a very fine introduction to psychological and educational statistics.

James A. Walsh
Iowa State University

Statistical Reasoning and Procedures by Robert B. Clarke, Arthur P. Coladarci, and John Caffrey. Columbus, Ohio: Charles E. Merrill, Inc., 1965. Pp. iv + 390. $7.95.

The authors of this recent Merrill publication perceive three levels of statistical skill: (a) the level at which the student can perform the clerical-computational operations needed in data summary and analysis, (b) the level at which there is sufficient understanding of statistical reasoning for the individual to be able to interpret statistical results as reported in the literature and to carry out research with the aid of a statistician, and (c) the level of specialization. The book has been prepared to help the student acquire a degree of sophistication at the second level, i.e., where he can communicate with the specialist and where he can know when he needs specialist assistance.

The content of the book and the sequence and format of presentation are staidly "traditional." An appeal has been to the non-mathematically inclined student so that the approach is primarily verbal. Mathematical notation and symbols have not been avoided, but they have been kept to a minimum. An appeal has also been made, however, to the instructor who adheres to an attempt to impart a thorough understanding of the "bread-and-butter" statistical techniques in contrast to the instructor who—in an introductory course—attempts to impart a feeling for the limits and the potentials of statistics. For a first course in statistics, there is more than adequate detail: indeed, there is sufficient detail to make this volume quite a useful reference at an elementary level.

The worthwhileness of the purpose of the text cannot be questioned. The "middle" level of statistical sophictication is needed increasingly amongst those entering the major professions in modern society. There is no scarcity of texts covering the content of the present book, however, so that it does seem reasonable to question the need for yet another in the area. The authors make no claim for having contributed anything unique or new. The successful payoff for their efforts, therefore, seems to depend in great measure on whether their particular style and presentation will prove more attractive than competitive texts.

Happily for the prospects for the book, the author's prose is highly readable. Explanations are patient without being tedious. The desire of every author for clarity has been achieved *par excellence* in this instance. Numerous diagrams—each worth at least a thousand words—have contributed considerably towards understanding.

The authors have not been helped in their presentation by the printed format. For example, the numerous worked examples (all from psychology or education) are carried through in cursive style in the printing, rather than placing one step to a line, which would

have helped comprehension. Admittedly, this may have been done to save on costs, but it detracts from the utility of the exposition. Similarly, nothing has been done to draw attention to summarizing materials. The over-all effect is that of a somewhat antique production. The confusion of X and chi in the appended table for chi-square appears to be the only obvious symbolic error in the text. Little need be said to the initiated user of statistics texts about the content of the book, presumably, once the label "traditional" has been affixed. There is the "usual" procession from measures of central tendency, through variability, the normal curve, correlation and regression, probability, hypothesis testing, t, statistical inference, and chi-square. The book stops short of the F-test, analysis of variance, and non-parametric techniques.

There are rare occasions for impatience. The first chapter, which includes a brief introduction and an even briefer summary of arithmetic operations and algebraic conventions, is a poor indication of what is to follow. To try to give a reader "some grasp of certain arithmetic and algebraic conventions" in five pages is a forlorn hope if he has had no prior acquaintance with these rules and redundant if he has. Again, the "interested student" is told from time to time in the text that he can "consult one of the many available treatments" of a given topic—yet there is no bibliography, no suggested additional or supplemental readings, no references. Each chapter ends with a number of excellent problems, but no answers are provided, so that there is no chance for the student to obtain immediate feedback as to his success or otherwise, without supplemental effort on the part of the instructor.

On the whole, however, the organization and substance of the content bear little criticism. Perhaps more use could have been made of titles and headings in drawing attention to important sections of the discourse and in improving the instructional merit of the text. Any such criticism or minor irritation is heavily outweighed by the clarity, precision, and thoroughness of the development.

This, then, is a book that will not command attention for innovation or uniqueness. It may find a market amongst present users of "traditionally-oriented" texts who have become disenchanted, or bored with, their current choice. It would certainly merit the consideration of persons interested in a well-written, careful exposition of the basic statistical techniques. It should also find a useful place as a modern reference work in the collection of any instructor of a first course in statistics.

Peter A. Taylor
University of Illinois

Statistics in Education and Psychology by Merle W. Tate. New York: The Macmillan Company, 1965. Pp. ix + 355.

Occasionally a book is published for the introductory college course in educational/psychological statistics which is basic without being fatuous. Such a book does not assume mathematical training beyond elementary algebra, but it does assume intelligence and perseverance. Its author recognizes that although the amount of mathematical training in high school and early college has increased somewhat in recent years, many of the students who enroll in required elementary statistics courses seem to defy this trend. They may have been attracted to major in one of the social sciences because they are more verbal than quantitative in ability and incorrectly perceive social science as non-quantitative. But there is a correlation between mathematical sophistication and logical reasoning ability, and many have observed that those students who have problems with the algebra in an elementary statistics course also have difficulty understanding the non-algebraic concepts.

Thus, the book which is the subject of this review is elementary in that it assumes little mathematical background, but it will not be elementary to the mathemaphobe or to one who has an antipathy toward a succinct, logical presentation of essentially mathematical material. The book is a condensed revision of Professor Tate's 1955 *Statistics in Education*. It gives the deceptive appearance of being a small book (355 pages), but within its cover is a thorough, tightly-packed presentation of statistics from frequency distribution through *t* test. The *F* test is not treated, although its chi-square counterpart is.

The author, Merle W. Tate, Professor in the Graduate School of Education at the University of Pennsylvania, is obviously a careful, accomplished writer and a serious scholar of statistics. The reviewer is quite pleased with this book and enthusiastically recommends it for courses in elementary statistics. It is not just another statistics book but an excellent presentation of much material. The instructor will have to work, however, it he plans to use this book, since the author wastes few words, and reiteration, exemplification, and evaluation by the instructor will be necessary.

The format of the book is a preface, twelve chapters, references, an appendix, a list of 10 tables, answers to selected exercises, and an index. The chapters are of unequal lengths, chapters five and eleven being relatively shorter and chapters seven, eight, and ten being significantly longer than average. The disproportionate amount of space devoted to chapters seven and eight reflects the rather unusual elaboration, for a beginning statistics book, of elementary test theory. Chapter seven might well have been broken into two chapters.

There are some differences between the organization, language,

and formulas of this text and what one is accustomed to in elementary statistics books. For example, a separate chapter is devoted to score transformations, whereas this material is usually covered in chapters on variability and the normal curve. Another example is the discussion of percentiles in the chapter on variability instead of in a separate chapter. Other differences are in the discussion of skewness and kurtosis, which is placed in the chapter on variability, and the use of the term "statistical series" as a central concept. Finally, setting $s = \sqrt{\Sigma x^2/N}$ instead of $s = \sqrt{\Sigma x^2/(N-1)}$ in Chapter IV makes for a certain conflict with traditional presentation when the standard error of the mean is defined as $s_M = s/\sqrt{N-1}$ in Chapter 10.

The book contains numerous formulas, definitions, and procedure descriptions, and it should be useful as a reference book as well as a textbook. There are a number of exercises at the end of each chapter, in which the student is required to reason and calculate. The proofs of certain propositions in the text are also left as exercises for the student. Significant of the carefulness with which the book was prepared is the fact that the reviewer detected only one typographical error in the whole book. On page 219, the word in the heading should not be "PROBABABILITY"!

The twelve chapters of the text are: I. Introduction; II. Organization and Presentation of Statistical Data; III. Characteristics of Statistical Series, Central Tendency; IV. Characteristics of Statistical Series, Variability; V. Transformation of Scores; VI. The Normal Curve; VII. Correlation and Regression; VIII. Reliability and Validity of Statistical Evidence; IX. Statistical Inference; X. the Normal Sampling Distribution; XI. The t Sampling Distribution; XII. The χ^2 Sampling Distribution.

Chapter I is a good introductory chapter on what statistics is and how it should be applied. It discusses the meanings of statistics, uses of statistics in data reduction and inference, and sampling. Chapter II is a short, general chapter on rules for grouping data in the form of a frequency distribution and constructing histograms and polygons. Brief descriptions of skewness, kurtosis, and the normal curve are also given. Chapter III gives a clear explanation of averages—mode, median, and arithmetic mean—and the conditions under which each is applicable as a measure of central tendency. Chapter IV on variability, as noted above, discusses the use and computation of percentiles, skewness, and kurtosis as well as the variance and standard deviation. The computing formula $s = 1/N\sqrt{N\Sigma X^2 - (\Sigma X)^2}$ for the standard deviation is a useful one not commonly seen. Chapter V discusses percentile rank and standard score transformations of raw scores. Chapter VI introduces the student to the normal curve, much attention being paid to normal transformations of data in either qualitative or quantitative

categories. Chapter VII is an excellent, if lengthy, chapter on correlation and prediction, although it will require a great deal of thoughtful study by the unsophisticated. The product-moment, rank order, biserial, fourfold, partial, and multiple correlations are all considered in this chapter. Chapter VIII contains much useful information on the theory and treatment of test scores. This chapter will also require careful study. Chapter IX is a brief discussion of statistical inference and involves more words and fewer statistical symbols than the majority of the chapters in the book. Chapter X is a comprehensive chapter on elementary tests of significance employing the normal curve tables. Noteworthy in this chapter is Table 10.3, which lists the standard errors of 11 different statistics. Incidentally, the square root in the denominator of the first formula on page 250 should be N, not $N-1$. Chapter XI is a brief, but thorough, chapter on *t tests*—small sample tests of significance of the difference between means and between correlation coefficients. The final chapter discusses various uses of the χ^2 distribution, although the author appears to prefer g statistics to the χ^2 goodness of fit for testing the assumpion of normality in a frequency distribution.

LEWIS R. AIKEN, JR.
Trinity University

Elementary Applied Statistics by Linton C. Freeman. New York: John Wiley and Sons, 1965. Pp. viii + 298. $6.95.

The number of new elementary statistics texts appearing on the market is quite astounding. Equally astounding is the variety and individuality of presentation of the same basic material. What is perhaps a little encouraging in this otherwise overwhelming situation is that each new book seems to have optimal utility for some particular class of student, so that somewhere amongst the welter of new textbooks each student should be able to find a book very much suited to his own individual needs.

Freeman's new book appears to be oriented towards the student who is primarily interested in *applying* the basic statistical techniques, in having some understanding of the reasoning behind statistical usage, rather than an appreciation of the theoretical development of the statistics. As the author is quick to point out, it is neither a reference work nor a "technical essay" in statistics. To categorize the text as a "cookbook" would be unkind; yet in general strategy, its development occasionally comes close to being just that.

Since the author chooses to define *science* as the "study of relationships among variables," the text on the one hand places a heavy emphasis on associational techniques. Some very interesting juxtapositions and relationships result from this approach, with a high

potential for student insight. On the other hand, there is a strong emphasis on levels of measurement (following Steven's categorization), and most of the presentation is organized around relationships within and between the various levels. The two-track developmental sequence—the nature of the problem and the form of the data—has been employed to produce a readable, often attractive, text.

The book is divided into four sections: a section concerned with the background, or "entry behavior" needed for the study of statistics; a section about summarizing distributions on a single variable; one on describing the association between two variables; and a final section on statistical inference.

The first section—the background section—consists of three chapters which attempt to cover a great deal of ground in a few pages. Chapter one is devoted to the definition and illustration of various levels of measurement, despite its rather misleading title, "The Subject Matter of Statistics." Interestingly, only nominal, ordinal, and interval scales are considered. The distinctions between the scales are concisely and clearly presented, each scale being introduced by a brief definition, with further exemplification. Unfortunately, the positioning of the chapter in the general developmental sequence leaves the concept of scales of measurement somewhat dangling, and for the moment unrelated to the primary purpose of the book. Chapter two indicates very briefly some of the major uses (and misuses) of statistics, and distinguishes between descriptive and inferential statistics. The final chapter in the introductory section contains the content that so often seems necessary for an elementary course; yet, to anyone with any sophistication in mathematics, it seems too hopelessly sketchy to be of much use. In some eleven pages, chapter three attempts to outline "the basic arithmetic and symbolic skills that are required." There is about a page on summation notation and another two pages on the "rules" for the use of sigma; arithmetic is "reviewed" in two and a half pages; and the remaining four or five pages are devoted to parenthetical gems on how to read the square root table in the back of the book, how to round numbers, and how to compute proportions and percentages.

Granted that a large proportion of students taking elementary statistics in the social sciences are not numerically adept, it is extremely doubtful that this kind of chapter can satisfy the need for added algebraic competence. If one is intent on buying a text in statistics, the appropriateness of a hodge-podge of arithmetic and algebraic "facts" that should presumably form part of the prerequisite student behaviors must be seriously questioned. Statistics is, after all, a branch of applied mathematics. As such, it is of doubtful value to try to disguise this fact and to bend over backwards in

order to avoid the use of anything that might be considered "mathematics." Chapter three in the present book does little to reassure or to enlighten the unsophisticated. It may better have been either omitted or considerably expanded.

The second section of the book is concerned with summarizing distributions on a single variable. The section begins with an overview which introduces a number of key terms and briefly reviews graphical techniques in a non-quantitative fashion. The continued protection of the reader from the bogey of actual numbers seems unnecessary: for example, it prevents the author from devoting more than eleven lines to frequency polygons and histograms. The reader is also saved the mental effort of struggling with such an esoteric term as "kurtosis." "Peakedness" is regarded as sufficiently taxing.

Having survived thus far, however, the reader should find remaining chapters in the book to be considerably more inspiring. Continuing with the second section, chapter four discusses methods of summarizing distributions on nominal scales. The mode and variation ratios are defined and illustrated with excellent examples, and useful notes are provided indicating alternative equivalent statistics (for example, the "uncertainty measure" and Mueller and Schuessler's Index of Qualitative Variation). Chapter five presents definitions and illustrations for computing the median and decile range for ungrouped data while chapter six does the same for the mean and standard deviation. Although there is only one illustrative example for each technique, it has been well chosen and well expounded. In each case, the raw data for the example have been culled from the literature, which adds to the readability. Again, and throughout the book, reference to alternative statistics is most valuable.

Section C in the text is concerned with describing the association between two variables. The section goes through the six combinations of the three levels of measurement two at a time—sometimes a little glibly, but more often with considerable illumination. There is always something exciting about finding fresh material in an elementary text; and in this third section, the book provides a number of unusual encounters. For example, in considering the association between nominal scales, we meet Guttman's Coefficient of Predictability and mention of Tschuprow's T—surely an interesting innovation. Again, at the level of ordinal measurement, Goodman and Kruskal's Coefficient is presented in detail, while *tau* and *rho* are dispensed with in a total of eight lines—and both without formulation or exemplification. Person's r receives about a dozen pages' attention and is treated clearly and fully. For describing the association between a nominal and an ordinal scale, a modification of the Wilcoxon test is proposed (the author's own extention); for the

association between a nominal and interval scale the correlation ratio is recommended; while passing mention of the biserial coefficients and "point-multiserial correlation" round out an extensive exposition of Jaspen's coefficient of multiserial correlation" as the appropriate index for describing the association between an ordinal and an interval scale. The author's choice of indices has in each case been fully and clearly presented, and an example worked through. One might query the wisdom of the choice—or perhaps, better, the wisdom of the omission—in a few instances, but one of the major attractions of the book lies in its proposed use of some of these lesser-known indices. It is here that the avoidance of mathematical derivation may, however, be of maximum disadvantage. Without reference to the original source (and in most cases, none is given in the text) it may be difficult for a student to generalize beyond the specific example given. Without in any way trying to detract from a refreshingly innovatory approach, there is here a slight suggestion of that "cookbook" strategy mentioned earlier.

The fourth section of the book is an attempt to outline the rudiments of statistical inference. The section begins with a chapter which is concerned primarily with defining terms and illustrating the processes involved in statistical inference. The discussion, which is brisk and readable, follows a traditional pattern. The thirteenth chapter deals with testing hypotheses about two ordinal scales—in effect, testing the significance of the Goodman-Kruskal coefficient. *Rho* and *tau* receive only five lines this time. Testing hypotheses about two interval scales is similarly almost exclusively concerned with testing for the significance of r. Chapters 15-17 are concerned with the testing of hypotheses of the combination of scales: nominal-ordinal, nominal-interval, and ordinal-interval. The U-test and F come in for major attention, with very brief references being made to the runs test, median test, and—oddly—the t-test. Chapter eighteen returns to the problem of testing hypotheses about two nominal scales and discusses chi square in some detail. The presentation is consistently lucid, and one must comment again on the excellence of the illustrative examples. Some of the emphases, to the total exclusion of more traditional techniques, might leave a number of gaps for the student who is confronted with the task of reading and interpreting research literature. While bearing the deficiencies in mind, what is included in the text has been well done.

The text comes to a rather sudden end after the fourth section. There is a brief bibliography, primarily of an eclectic nature. Certainly the student who wished to pursue a specific topic would get little guidance from the list provided. More specific references at the end of each chapter would have been beneficial. A list of statistical symbols and the statistical tables referred to in the text,

together with a selection of data tables for use in problems, complete the appendix. A few problems (never more than ten) associated with each chapter and the answers to the odd-numbered problems also appear at the end of the book. The problems are simple, requiring minimal arithmetical manipulation, and are specifically related to the textual material. A brief index completes the book.

In making an over-all evaluation of the book, a return must be made to the title, *Elementary Applied Statistics*. It is difficult to decide at what point the book would find most use. It appears to be too elementary for a graduate course, yet some of the techniques it provides are unlikely to be utilized at an early stage in a behavioral science program. In many ways, the book appears to be a useful supplement to a more traditionally-oriented text. It makes a unique contribution to the flood of elementary texts in introducing a number of less well-known but powerful tests. Its style is highly readable and the illustrations are excellent. Therefore, it should be capable of rewarding individual study. And it is in this sense that the book is seen as being potentially valuable as supplementary reading. Its over-avoidance of mathematics, the lack of adequate references, and the relatively simple problems reduce its appropriateness as a primary text. Superficially, it would also appear to restrict teaching very much to the direct content—for those who dislike this strategy, some discomfort may be engendered. Again, this could be avoided were the book to be used in a supplementary role.

Elementary Applied Statistics must be commended for its fresh approach and for its expository style. It would deserve the attention of instructors of elementary statistics courses as valuable supplementary reading.

PETER A. TAYLOR
University of Illinois

Foundations of Behavioral Research: Educational and Psychological Inquiry by Fred N. Kerlinger. New York: Holt, Rinehart and Winston, Inc., 1964. Pp. xix + 739.

This compendious volume on behavioral science methodology should foster the growing sophistication and efficiency of students looking forward to conducting psychological, sociological, and educational research. It may be compared favorably with such previous methodological books as Underwood's *Psychological Research* and Festinger and Katz's *Research Methods in the Behavioral Sciences*. It is a far more comprehensive and difficult treatment of the material than the recent introductory text by Scott and Wertheimer. This book is suitable as a text or as a reference for upper division college and graduate students; it would be adaptable for a two

semester course on methodology when supplemented with additional references; it would provide valuable supplementary reading for courses on statistics and experimental psychology.

Kerlinger states that the major purpose for the book is to help students understand the fundamentals of the scientific approach to problem solving. The author outlines the scope, objectives, and subject matter as follows: (1) It is a treatise on scientific research. (2) It concentrates on helping the student to grasp the difficult relations between a research problem and "the design and methodology of its solution." It is not a methods book, but refers the student to references on the specific methods considered. Throughout the book there are detailed discussions on (a) stating research problems, especially on developing testable hypotheses; (b) selecting an appropriate and powerful design for planning for observations and analyzing the data; and (c) pointing out the logical relations between the problem under investigation and the research method used. (3) The content and the explication of research procedures are linked together through the notions of set, relation, and variance. These ideas, combined with the fundamentals of set theory, probability theory, statistical design and analysis, and measurement, are used to integrate the diverse content of research activity into a systematic and meaningful whole. (4) The book concentrates on educational and psychological problems, especially the psychological aspects of educational research problems.

The appropriateness of the methodology to the problem under investigation is based upon careful consideration of (a) a choice of research design, (b) methods of observation, (c) methods of measurement, and (d) types of analysis from which interpretations are derived. In other words, the design, methods of observation and measurement, and the statistical analysis must be appropriate to the research problem and should be chosen and combined in a way so as to obtain the maximum amount of information from the data. Significant research stems from the complex chain of theory, deductions from theory in the form of hypotheses, design, measurement, and analysis. The book is divided into eight main parts, each one treating an important aspect of the logic and procedures of research.

Part One discusses the salient characteristics of the scientific approach, "Scientific research is systematic, controlled, empirical and critical investigation of hypothetical propositions about the presumed relations among natural phenomena." Definitions and examples of different types of hypotheses, constructs, and variables are presented in detail. The notion of randomness and sampling from sets is included in the introductory section.

Part Two briefly treats the main principles and concepts of set theory as a way of giving the student an appreciation of the formal

language underlying the important operations of scientific methodology. The consideration of variation as a universal characteristic of natural phenomena leads to the definitions of different kinds of variance, including population and sample variances, systematic variance, experimental or between-groups variance, and error variance. Examples of systematic and error variance lead intuitively to the notion of statistically significant differences. The operations that can be performed with variances are summarized under the categories of components of variance, covariance, and common-factor variance.

Part Three, comprising approximately one fifth of the book, summarizes the basic concepts of probability theory and statistical inference, emphasizing the role of statistical hypothesis testing and analysis of variance in behavioral science research. The procedures outlined in this part give primary consideration to the rationale for determining significant statistical differences, although the author stresses in a single sentence the need for consideration of estimation, interval estimation, confidence intervals, and exact probability methods, since they give the research worker more information about the relations found in the data. Several paragraphs are devoted to an interesting discussion of the relation of substantive and null hypotheses. The author gives somewhat limited attention to the meaning and importance of interaction in analysis of variance designs.

Part Four, Designs of Research, offers a set of principles, which every student of statistical and research design would benefit from reading. Research design is seen as having two basic purposes: (a) to provide answers to research questions, and (b) to control variance. Research designs, generally, should maximize experimental variance and minimize error variance. Kerlinger discusses six steps by which these two goals may be reached. Chapter Sixteen, "Poor Designs," is one of the highlights of the book, in the reviewer's opinion.

Part Five contains an extensive discussion of the characteristics, criteria, advantages and disadvantages of various types of research, ranging from laboratory experiments, field experiments, field studies, and surveys. He differentiates between experimental research and ex post facto research on the basis of the kinds of controls, sampling, and interpretations of data characteristic of each. He points out that behavioral research frequently must be content with using ex post facto methods, but the precision and scientific validity in the use of such methods may be increased if research workers would concentrate upon a set of alternative hypotheses and then treat the results and the interpretations made therefrom cautiously.

Part Six deals with the foundations of measurement, including levels of measurement and scales. Then reliability theory is intro-

duced as an extension of the notions of set, relations, and variance. Reliability is defined as "the proportion of the true variance to the total obtained variance of the data yielded by the measuring instrument." Conversely, "reliability is the proportion of error variance to the total obtained variance of the data yielded by a measuring instrument subtracted from 1.00, the index 1.00 indicating perfect reliability." The statement that reliability is a necessary but not sufficient condition of the value of research results, leads to the consideration of types of validity. Validity is defined as the proportion of the total variance of a measure that is a common factor variance. Kerlinger summarizes this part of the book as follows: "Poor measurement can invalidate any scientific investigation. Most of the criticisms of psychological and educational measurement . . . center on validity. . . . Achieving reliability is to a large extent a technical matter. Validity, however, is much more than technique. It bores into the essence of science itself. It also bores into philosophy. Construct validity, particularly, since it is concerned with the nature of 'reality' and the nature of the properties being measured, is heavily philosophical."

Part Seven, Methods of Observation and Data Collection, classifies and discusses a heterogeneous assortment of methods by the degree of their directness. Among the methods reviewed are interviews and questionnaires, objective tests and scales, critical incident technique, the problems involved in defining units of behavior, and the scales appropriate to various units, rating scales, projective methods, and the use of available relevant materials and the problems involved in content analysis, sociometric methods, semantic differential, and Q Methodology.

Part Eight briefly describes the processes of inference involved in data analysis and interpretation, and then presents still further procedural methodology, including relevant methods from set theory and statistics. A short section on the interpretation of research data succintly summarizes many of the research precepts which Kerlinger wishes students to grasp. This part of the book includes chapters on the analysis of crossbreaks (introduced in the first part of the book) and factor analysis. The use of factor analysis to test hypotheses about the relations among variables in addition to that of exploring and identifying variables is described.

The book concludes with three short appendices, the first one outlining some main points on report writing and pertinent references, the second one reviewing the differences between so called historical and methodological research, and the last one very briefly indicating the value of electronic digital computers in conducting behavioral research.

The book contains both subject and author indices, which are fairly accurate. Several typographical errors were evident. In sev-

eral contexts the reviewer would have welcomed a more extensive and varied use of illustrative research reports. This book, to which the author has given a great deal of thought, deserves a place in collegiate and students' libraries.

CHERRY ANN CLARK
The Meyers Clinic, Los Angeles

Psychological Tests and Personnel Decisions (Second Edition) by Lee J. Cronbach and Goldine C. Gleser. Urbana, Illinois: University of Illinois Press, 1965. Pp. viii + 347. $7.95.

The potential impact of decision theory on testing theory and practice is great, and in their 1957 edition the authors spelled out many of the important changes in viewpoint and practice likely to take place when testing is regarded as a part of a decision-making process and not as an end in itself. A check of some past issues of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT indicates that this new revised edition of this still basic volume may be urgently needed for another reason in addition to that of up-dating: few references to the original edition were found, although references to Cronbach's *Essentials of Psychological Testing* were frequent. With a book as important as this one it would be very desirable if the ratios of mention were reversed, with familiarity with "EPT" taken for granted, but "PTPD" cited frequently as a stimulus for work in the broader context of decision theory.

In this Second Edition, the authors have added a valuable thirty-page summary of recent work in decision theory related to selection problems and to the use of tests. They reported that "while the ingredients represented in our survey might conceivably be combined into some great new insight, we cannot yet report that we or others have made any such advance over our first edition position. Rather, we see recent writings as pushing ahead toward more definite applications of certain principles in actual situations, and in another vein, toward extremely abstract analyses of various problems that move farther and farther from application."

Included in a supplement are representative and valuable papers by other authors, some new, some reprinted, illustrating possible applications of decision theory to selection problems as well as raising (inevitably) the methodological and practical issues involved. Although much of the treatment is mathematical, the generalist in measurement is given an instructive view into the boiling cauldron of flickering assumptions and mathematics too hot to handle except for those with much training: he is likely to be well convinced of the potency of the decision theory approach, and grateful that the fires of criticism will sear away the dross. However, he may have his present commitment to classical concepts in testing

(as a matter of sound, more sound, and most sound measuring instruments) somewhat scalded. The wound will heal. For quite some time, the new tissue may not be so strong as it was before, but it will be more sensitive to the realities of the world in which the tester lives. Greater perception of how testing fits into the real world will come, it seems, when one's toughened measurement callouses are replaced by receptors more sensitive to the storms and hail of a difficult world where tests are used to make decisions in the dim light of uncertain utilities.

One should be sensitive, the authors urge, to the fact that a test with high validity (such as an aptitude test predicting college grades) may not be so useful as it seems, since the test is competing with high school grades as a predictor when this (free) information overlaps considerably with test score predictions. The authors do not necessarily mean that such aptitude tests are of diminished value; they do indicate that "The characteristics of the specific decision determine 'what the test is worth'," and that "a test of validity .20 in one situation may be more beneficial than a test of validity .60 in another."

Decision theory leads them to important discussions of alternative strategies in single-stage, two-stage and multiple-stage decision making (e.g., in selection problems) and to point up the "'bandwidth-fidelity" dilemma. It is suggested that for some purposes a broad band technique such as an interview, touching on many aspects of the interviewee, may provide more over-all utility even though each of the many items of information it provides is less reliable than a single score from a very reliable ability test. The authors are of course not arguing against sound measurement practices. They are arguing that measurement is for a purpose the utilities of which should determine what optimum degree of accuracy in measurement is most useful.

The authors also point out the importance of adaptive treatment as an alternative to better selection. "The job simplification expert and the human engineer seek to fit the job to unselected men. The greater their success the less the value of selection. The tester has failed to realize that he is competing with the treatment simplifier."

Emphases such as those abstracted above have of course been explicitly and implicitly foreshadowed in the literature for years, and the authors do cite such previous work. But for the person interested in testing, in personnel problems, and in learning at this time something of what graduate students in testing will probably be routinely required to learn in 1970, this book is a must. For those who have already acquired the 1957 edition, the supplements and the guide to the relevant literature from 1955 to 1963 should make the book worth the re-investment. As an up-to-date discussion of the anticipated huge impact of decision theory on personnel prob-

lems in general and testing practice in particular, this book is strongly recommended.

T. B. Sprecher
Educational Testing Service

Studies in Item Analysis and Prediction by Herbert Solomon (Editor). Stanford, California: Stanford University Press, 1961. Pp. xi + 310. $8.75.

Constituting an important addition to the Stanford Mathematical Studies in the Social Sciences *Studies in Item Analysis and Prediction* is a collection of 19 papers by ten renown contributors to the professional literature in mathematical statistics. Edited by Herbert Solomon who for many years has been quite interested in the application of statistical methodology to educational and psychological measurement, this volume is primarily concerned with the development of mathematical models that describe optimal designs and approaches to problems of prediction and classification. In particular emphasis is placed upon the formulation of optimal test designs in terms of which items with what are deemed optimal properties may be selected, although the computational difficulties associated with use of the models have placed serious limitations upon the practical length that a given test can assume. Nevertheless, one may anticipate that progress both in mathematical decision theory and in computer technology may eventually make feasible the application of several of the models proposed. Certainly this volume lays an impressive and basic foundation on which additional creative research efforts in test theory and in test analysis can be expanded.

Preceding the three major divisions of the volume is an introduction by Herbert Solomon—an overview of the entire book. In a highly lucid style Solomon not only has outlined a rationale underlying test theory, but also has provided a succinct statement of the scope and significance of the contributions of each chapter often in its relation to the contents of other chapters of the book. In fact, his discussion of each paper, which allows the psychologist with limited mathematical experience to gain the gist of each contribution, serves to clarify and to amplify in a meaningful way what has sometimes been only implied or somewhat obscurely developed in several of the papers.

In Part I the first seven chapters are concerned with item selection procedures within the framework of a multivariate normal structure. The first three chapters form a logical unit in which problems of reliability and validity (including the attenuation paradox) are considered. In terms of the assumptions of a unidimensional trait, dichotomous scoring of items, the description of item characteristics as normal ogives, and a normal distribution of the

measured trait in the examinee group, a theory of test design is systematically developed. The remaining chapters in Part I are largely concerned with the problem of optimal selection of test items without the need of determining the almost infinite number of possible multiple correlation coefficients associated with almost countless composites of test items.

The development of nonparametric probabilistic models for handling dichotomous systems of data and for classifying students, especially in relation to modern statistical decision theory, furnishes the main emphasis underlying Chapters 8 through 14 in Part II. In Part III the inherently difficult *W* classification statistic devised by T. W. Anderson and the determination of its approximate sampling distribution are treated at length in the last five chapters. The mathematical complexities in the chapters of Part III are formidable almost beyond belief.

What the impact of this definitive volume will be upon future developments of test theory remains to be seen. Until greater numbers of psychologists receive the requisite training in higher mathematics which one would estimate to be at a level considerably in excess of an M. A. degree in mathematics additional theoretical developments will be somewhat limited. Perhaps more serious than the absence of psychologists with mathematical training approximating the doctoral level is the lack of a necessary group of individuals with intermediate levels of mathematical maturity and with substantial interests in measurement and evaluation who can translate these esoteric theoretical developments into the day-to-day programs of personnel selection and classification. Unfortunately, it is not unlikely that at least twenty or thirty years will expire before any substantial or noteworthy proportion of the theoretical developments of this truly significant book will become operational in local, state, or national testing programs.

William B. Michael
University of California, Santa Barbara

Spatial Ability: Its Educational and Social Significance by I. Macfarlane Smith. San Diego: Robert R. Knapp, 1964. Pp. 408.

American education appears to be faced by some of the same problems that are faced by education in Great Britain. This book proposes an approach to one of these problems that should be of interest to educators and psychologists in both countries. The book begins with a discussion of the shortage of scientific manpower and then proposes that this could be alleviated if technical schools would give consideration to spatial ability in their selection of students. The disregard of spatial ability by schools results not only in a shortage of trained technical manpower, but also in the presence in the schools of a group of children who possess talents they

are not called upon to use. The rest of the first half of the book is concerned with factor-analytic studies of spatial ability. In the second half of the book there are discussions of the relationship of spatial ability and abstract thinking, temperament, motor perseveration, critical fusion frequency, alpha rhythm, and the reticular brain stem formation. The appendix includes a number of short biographical sketches of famous persons whom Macfarlane Smith has found to illustrate his conception of spatial ability or the lack of it.

According to Macfarlane Smith, spatial ability is a talent of crucial importance to the prospective scientist or technician. It is a talent which is neither recognized nor exercised appropriately in school. This reviewer judges that he had a good case. The book, however, is more hortatory than dialectic—the author cites study after study and case after case in a way that seems more likely to importune than to clarify the reader's mind. Since this book is obviously addressed to educators as a primary audience, this approach may be an appropriate strategy. However, to the scientist this style may be somewhat less appropriate. Smith has accumulated a great deal of information pertinent to his topic from a wide variety of studies. A more efficient organization of all of this material would have presented problems.

Having had some experience and previous interest in this field, this reviewer read the book primarily in the hope of at last learning what "spatial ability" was. In this respect the book was, perhaps inevitably, a disappointment. Smith does not make this a central issue. On page 55 we find this statement: ". . . the special aptitude (spatial ability) . . .would be manifested in an ability to perceive and reproduce shapes correctlty, i.e. with their dimensions and their relations in due proportion." This is about as complete a definition of the trait as is offered in the book at any single point. However, at various places there are additional qualifying remarks, such as: ". . . spatial ability is quite as intellectual as verbal or numerical abilities" (page 36); or ". . . there is a correspondence between . . . spatial ability . . . and a 'fixative' or 'concentrative' mode of attention . . . a tendency to attend to a configuration as a whole . . . masculine attitudes . . . schizothmia . . . emotional stability" (Page 238). Although this discussion may represent a marked advance over previous definitions of spatial ability, it appears that there is still much to be learned.

This book is a clear demonstration that factor analysis is a very permissive technique—not only in the type and amount of rotations and other mathematical operations, but also in the psychological interpretation that may be deduced from the numerical tables. The psychological interpretations are subjective. Moreover, they are influenced both by one's knowledge of the nature of human nature and

by one's familiarity with tests as well as by one's attitude toward the meaning of tests. Smith's contribution is that he has brought together a very broad knowledge of tests with what appears to this reviewer to be a sensible view of the nature of man. Since among psychologists there is little consensus regarding the nature of man, Smith's interpretations will not be convincing to all psychologists. Fortunately, the data cited by Smith are not restricted either to studies of cognitive tests or to the results of factor analyses. He includes chapters on temperament and on various physiological studies. This breadth of view is one of the main virtues of the book.

This book may be commended to the attention of persons other than just those interested in spatial ability. The problems and challenges faced by those interested in spatial ability twenty years ago have something in common with the problems and challenges faced by the students of "creativity" today. And the case that can be made for the importance of spatial ability to education can, perhaps, be made for other talents as well. Macfarlane Smith proposes that education may be too exclusively verbal. Surely he is right, although psychologists may have yet to discover how to educate the verbally inept. Anyone who wishes to broaden the base of education may find some interest and value in this book.

Charles T. Myers
Educational Testing Service
Princeton, New Jersey

The Structure and Measurement of Physical Fitness by Edwin A. Fleishman. Englewood Cliffs, N. J.: Printice-Hall, Inc. 1964. Pp. 207.

Examiner's Manual for the Basic Fitness Tests by Edwin A. Fleishman. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964. Pp. 60.

Historically, there have been alternating waves of public interest in the subject of physical fitness, with concern high during war or threat of war, and low during periods of comparative peace. Today the attainment of a high level of fitness is an international aim. With reference to the current concern, the United States (at the White House level) was first shocked into action by a July 1955 report (widely publicized by the press) which indicated that European children were more fit than their American counterparts. Though attempts to identify levels of fitness are far from new, dating back to the latter part of the nineteenth century, the seriousness of the contemporary situation has sparked innumerable efforts to devise adequate physical fitness tests.

Dr. Fleishman, currently Director of the Washington Offices of the American Institute for Research and a consultant to the Surgeon General of the Army, has recently published the results of a

series of studies prepared under contract with the Office of Naval Research. The investigations were centered at Yale University, where from 1957 to 1963, Fleishman was Director of the Human Skills Research Project and Professor of Psychology. The book and manual reviewed below contain the "lay" version of the results of Fleishmen's extensive technical investigations.

Fleishman and his associates attempted, with effectiveness, to isolate dimensions of physical proficiency. Starting with factors which through the years have been empirically deduced and, in some cases, previously identified by physical educators through factor analytic means, Fleishman located and/or postulated fourteen dimensions of proficiency. Each was then intensively studied by means of the statistical technique of factor analysis.

An attempt was first made to conceptualize a strength and endurance area. The parameters were found to include "dynamic strength" (the ability to exert muscular force repeatedly, or for long periods of time). This factor was seen to be the most general in the strength domain; "trunk strength" was found to be a specific secondary dynamic strength component. Also identified was "static strength" (commonly now called "isometric strength—the ability to exert force against heavy or immovable objects). A third parameter was termed "explosive strength" or "energy mobilization" (probably synonymous with the physical educator's term "power"—the expenditure of force).

Next were tackled performances which apparently emphasize chiefly speed, flexibility, balance, and coordination. Components identified were two kinds of flexibility—"extent flexibility" (the ability to stretch, flex, and extend as far as possible) and "dynamic flexibility" (flexibility in action); "speed of limb movement;" "gross body equilibrium;" balance of both the static and dynamic type which depend chiefly upon nonvisual cues, and balance which requires visual cues. In the tasks chosen to represent this speed-flexibility-balance-coordination area, explosive strength, previously identified, again emerged. Agility, however, was not identified as an independent factor.

Thus Fleishman identified seven factors of physical proficiency: extent and dynamic flexibility; explosive, dynamic, and static strength (with trunk strength as a secondary component); gross body coordination; and gross body equilibrium. It was hypothesized that cardiovascular-respiratory endurance should, on logical grounds, be added to this list.

In addition to describing the details of the analysis of tests for each of these categories, *The Structure and Measurement of Physical Fitness* includes general discussions of 'Abilities and Motor Skill," "What Do Physical Fitness Tests Measure?", specific information regarding Fleishman's subsequent work establishing na-

tional norms for the tests which he suggests as measures for each of the identified factors, developmental curves in relation to age and sex, and recommendations for a battery of basic fitness tests.

The manual presents, in concise form, test descriptions, administrative procedures, evidence of validity and reliability, and national norms for boys and girls, age 12–18, based on performance records of 20,000 young people in 45 cities across the country.

This work is monumental in scope, the most extensive and intensive effort yet attempted to isolate dimensions of proficiency. It confirms factors previously identified by other investigators; in addition, new factors have been isolated and some old ones redefined.

Nowhere is the larger concept of *Total Fitntess* mentioned (fitness, broadly conceived, involves the integrity of the total organism: social, mental, emotional, and spiritual aspects as well as physical). No doubt Fleishman's work will be criticized because of this lack. This reviewer feels, in Fleishman's defense, that he never intended to imply that the physical aspect of fitness constitutes its totality; the name of the book properly limits its scope.

Perhaps an important and ultimate value of Fleishman's efforts, though not his purpose, may be recognition that his data provide the most conclusive evidence yet advanced for belief in the extreme specificity of motor performance.

Because the publication of the results of this voluminous amount of investigation occurred only after a nationwide motor fitness testing program had already been not only initiated but also well advanced by the American Association for Health, Physical Education, and Recreation, Dr. Fleishman's recommendations may find their way into practice rather slowly. This possible lag is unfortunate. His battery of tests requires little or no equipment. Moreover, the tests can be readily administered and scored by almost anyone who gives them careful study.

AILEENE LOCKHART
University of Southern California

Evaluating Pupil Growth by J. Stanley Ahmann and Marvin D. Glock. Boston: Allyn and Bacon, 1963. Pp. 640.

The volume, *Evaluating Pupil Growth,* is intended to be used as a textbook for a one-semester course in Measurement and Evaluation for students who are preparing to be teachers in either the elementary or the secondary school. Evaluation is broadly interpreted to mean the appraisal of the degree to which the objectives of the educational program are reached as they affect pupil growth. Evaluation is assumed to include the use of all types of techniques and instruments, tests and non-tests.

This volume is the second edition of a textbook first published in 1959 by the same authors. From comment made in the preface one can assume that certain professors and students who used the first edition of this book made recommendations for improvements some of which were implemented in this second edition.

Evaluating Pupil Growth possesses some characteristics which make it unique as a textbook in this field. One of these unique characteristics is the arrangement of the chapters. Most textbooks in *Measurements* start with a chapter such as Measurement Today. Then follow chapters on the History of Measurement, Statistics, Validity and Reliability, Teacher-made Tests and Standardized Tests. This book postpones the Validity and Reliability chapters until after the chapter on the Teacher-made Tests. To be able to do this the authors do have to discuss these two topics, although briefly, before the chapter on Teacher-made Tests. They then devote an entire chapter to each of the topics, Validity and Reliability, in the final half of the book. The reviewer is not commenting adversely about this arrangement of major topics of the book; he is simply saying that it is unusual. Furthermore, a professor in teaching a course may take the chapters in any sequence he wishes.

Another unique characteristic of *Evaluating Pupil Growth* is the use of questions and exercises called problems placed appropriately throughout each chapter. To the reviewer this gives the book a real strength. Of course, the problems may be used as the teacher of the class dictates, but the student who reads and answers them as they apear during the course of reading the chapter will find them a distinct help in enhancing his understanding of the content.

A third unique feature of the book is the use by the authors of novel, well selected, interesting, narrative or descriptive, stories or discussions at the beginning of each chapter. These are all appropriate to some of the content of the chapter. The first chapter, for example, begins with a summer beach scene where the physical and psychological differences among the beach population first and then among people in general are described. The reading material then moves on to the differences among pupils in a school and finally to the need to evaluate these differences. Each chapter in a similar manner begins with illustrative material to stimulate interest and then leads into the topic of the chapter. As one reads the book, he is delighted with these little digressions and often thinks forward in anticipation of what anecdote or illustration will introduce him to the next topic.

Although it is not a unique feature of this book, it is written with a substantial number of well selected and appropriate figures and tables. Illustrative of these are: (1) A Pupil Profile Chart of the STEP achievement test, (2) A table of Educational Objectives in Communication from *Elementary School Objectives* by N.

C. Kearney,[1] and (3) A Table of Quartiles, Deciles, Percentiles, and Percentile Ranks Based upon a Distribution of 360 Raw Scores. One who is accustomed to teaching the area of Measurement and Evaluation knows how useful is each of these two tables and the figure. Many others of similar value are found in the book.

The strongest section of the volume is Part II dealing with the Informal Methods of Evaluating Achievement. In five chapters the authors in a most lucid manner show how needs are translated into objectives which in turn are evaluated as knowledge, understanding, procedures, and products with the teacher's tests and evaluation instruments. Good examples of types of tests are given, and the use of a wide variety of types is urged.

Each chapter concludes with a summary and with a well selected list of annotated Suggested Readings. Appendix A contains the material on statistics which is commonly found in measurement textbooks as a chapter of the book. This material is condensed into eleven pages. There is a very brief presentation containing one measure of central tendency, the mean; one measure of variability, the standard deviation; the product-moment coefficient of correlation; and three measures of relative performance, the quartile, decile, and percentile. If this material is all of the statistics learned by students in the course for which this book is intended, obviously it is a very shallow coverage.

Appendix B contains two and one-half pages of the names of publishers and distributors of standardized evaluation instruments. Appendices C, D, and E consist of eleven pages of tests with supplemental data of publisher name, year of publication, age or grade level intended, time required for administration, and a reference to the most recent review in Buros' *Mental Measurements Yearbooks*. Following the last appendix but before the index are fifteen pages of well selected references listed by chapter and in the order in which they were cited in the text.

One of the least satisfying chapters to read and presumably one of the most difficult to write is the one on Standardized Achievement Tests. As stated earlier the book is a general text in that it is written for students planning to teach in the elementary school as well as for those who plan to teach in the secondary school. Authors of measurement textbooks designed for both groups of teachers do have the problem of deciding how many achievement tests at each school level should be discussed and, since there are so many such tests published, which ones should be chosen for discussion or comment. Any tests selected for inclusion will automatically eliminate other good ones unless authors mention no

---

[1] Kearney, Nolan C., Elementary School Objectives. New York: Russell Sage Foundation, 1953, pp. 102–110.

specific standardized tests and say in effect, "Go read *Tests in Print*, or Appendices C, D, and E." But this approach is not helpful to the reader who knows practically nothing about the subject. One assumes, therefore, that the authors have selected the tests to be mentioned on some basis. Those which were included the reviewer thinks are good. But there are no specific achievement tests mentioned in many high school subjects such as Mathematics, History, Foreign Language, or Science except as they are included in test batteries. Similarly, there are many good elementary level achievement tests apparently overlooked.

The same situation obtains but to a lesser extent with respect to the aptitude tests and personal-social adjustment tests. Many of these such as the Stanford-Binet Intelligence Scale and the California Test of Mental Maturity measure intelligence of both elementary and secondary school age children, the age and grade levels of the pupils to be taught by the teachers to be using the book. But the Minnesota Paper Form Board Test and the Differential Aptitude Tests are for secondary level students and adults. Prospective teachers of elementary school children have difficulty generating an interest in studying about such tests.

In using this textbook with a class, the teacher can be satisfied with the scholarly and interesting presentation of a growing but rather well defined body of subject matter. The authors are thoroughly familiar with the topics discussed and with the best references related thereto. It was necessary for the authors to determine what should be the major topics of a course in Evaluating Pupil Growth. In making this decision it was necessary to truncate some sections in order to give complete treatment to others. This reviewer reacts favorably to their selection of topics to be included. He is neither adversely critical nor suprised at the decisions as to what topics should receive a brief treatment. The references listed in the book are for the use of those who desire more breadth or depth.

Glenn W. Durflinger
University of California, Santa Barbara

Test Norms: Their Use and Interpretation by Frank B. Womer. Washington D. C.: Association of Secondary-School Principals, National Education Association, 1965. Pp. 56. $2.00.

This small booklet has been written for the purpose of clarifying concepts and procedures relative to test norms and norming techniques. Clearly, it has been written for pupils, parents, teachers, administrators, and boards of education rather than for the test specialist or for those sophisticated in the intricacies of educational measurement. It is written in a practical "down to earth" manner. It answers many pertinent questions frequently posed by teachers.

The percentile norm is used as the basis for discussion with an occasional mention of other types of norms. Generally, the booklet assumes little or no previous knowledge of norms by the reader. It is written in a language which the typical layman would comprehend, such as referring to norms as a "yardstick." However, concepts such as standard deviation, correlation coefficient, and reliability coefficient not only are undefined but also are evidently assumed to be previously understood by the reader.

Not forsaking rigor at the expense of oversimplification is a problem which one experiences when writing in a technical area for lay consumption. Dr. Womer has met this problem estimably, with few exceptions. One rather notable exception is the treatment of test reliability. He indicates that reliability is synonymous with accuracy rather than consistency. Expressions such as "average (median)" only serve to confuse readers who assume average to be the mean. An admirable, but perhaps questionable, effort was accomplished in discussing norms without introducing the concept of validity, *per se*. Error of measurement was developed in terms of sampling error rather than in terms of a standard error of measurement concept.

A reader specifically interested in gaining a conceptual basis for such norms as stanines, grade placements, and standard scores would have to obtain such information from sources other than this publication. However, for interpretation of norms in general, their limitations, applications, strengths, and weaknesses, this publication is an important contribution to school staff members who utilize standardized test results. The present reviewer recommmends this booklet especially to school administrators and board of education members.

JERRY C. GARLOCK
Los Angeles County Schools Office

An Experimental Approach to Projective Techniques by Joseph Zubin, Leonard D. Eron, and Florence Shumer. New York: John Wiley and Sons, 1965. Pp. xix + 645. $13.50.

It should be said at the start that this volume is probably the most important one on projective techniques to appear in recent years, and perhaps in many years. It is directed to clinical psychologists (at least those who will listen) and to psychologists of the general-experimental persuasion. The former will probably damn the authors, and the latter will derive tremendous satisfaction for having decided on non-clinical research activities.

The discussion of projective techniques is limited to the Rorschach and Thematic Apperception Test (TAT), or to some variation of the two most important techniques. This volume is not a clinical manual to assist the psychologist employing projective techni-

ques to arrive at better interpretations of test material. Indeed, after reading this book, one may wonder whether there is really any basis in psychological fact for writing such a manual. These writers have done more than review validity and reliability studies. They have systematically pointed out methodological and theoretical flaws and have cited the sheer weight of studies with results unfavorable to the projective techniques. Although others, of course, have cited such difficulties on more than one occasion, these authors do it again with skill and brilliance.

In essence, the writers have pointed out that users and investigators of the Rorschach, particularly, have been divorced from the mainstream of psychometric and experimental psychology in both theory and method. The importance of the external criterion for validation is emphasized and the difficulties of the criterion problem in personality measurement are stressed. The most damaging blows to the projective body are twofold: separation of "projective" technique theory from Freud's formulation of "projection" and divorcing projective techniques from perception. The latter is the more serious of the two because it neatly cuts the thread to psychological theory and findings that have given respectability to the projectionist.

The two chapters on perception will be of more than passing interest to experimental psychologists. To the clinical psychologist, these two chapters are critical because it apparently dissects the link between personality and perception. Social psychologists working in the area of person-perception will be made to feel a bit uncomfortable, too. But for the Rorschacher, it will be dismaying to have the authors demonstrate that the Rorschach is not a perceptual task and that perception offers little if anything in the way of a rationale for the instruments. The TAT, which is discussed separately, comes out with a cleaner bill of health than the Rorschach, as far as its having a foot in the psychological mainstream.

Having criticized the naivetè of "projective theory" as well as the lack of objectivity in research and in application and having shorn projective psychology of its links to experimental psychology, one might expect the authors to argue for the abandonment of the projective concept and method. This the authors come close to doing, and one suspects that they would really like to see projective techniques quietly interred. Instead of arguing for the abolition of the Rorschach and the TAT, they have provided procedures for giving these devices some potential respectability as psychometric devices. These procedures include a series of rating scales, which would avoid many of the serious problems found when an investigator correlates Rorschach scores with an outside criterion. Presumably, these scales will face similar problems in determining reliability

and validity now facing the Rorschach and TAT as currently "scored."

What will be the impact of this book on practice and training in the projective techniques? The authors are under no illusion that their position is the final nail in the coffin of projective psychology or that projective methodologies will soon join (to use their examples) phlogiston, marasmus, and spontaneous generation—concepts abandoned by other sciences. Based on personal experiences, the reviewer can say that the clinical psychologist cannot read this book without developing a vague sense of uneasiness and apprehension. Is it possible that some will enter the tabernacle to sneer, but stay to pray?

PHILIP HIMELSTEIN
Texas Western College

The Rorschach Index of Repressive Style by Murray Levine and George Spivack. Springfield, Illinois: Charles C. Thomas, 1964. Pp. vii 164.

This small volume is an impressive example of a projective test manual and a provocative work on construct validation. Levine and Spivack, two research psychologists, set for themselves the task of developing "a measure of individual differences in the proneness to repress," by studying various ideational qualities reflected in Rorschach responses. The book presents the psychological "engineering," which led to the development of a rationale and a formalized system of scoring for the Rorschach Index of Repressive Style (RIRS).

The authors acknowledge their indebtedness to the writings of a number of Rorschach workers for their techniques of clinical analysis of significant stimulus qualities. In the initial stages of investigation, they recognized that the RIRS does not measure accomplished acts of repression, but it seems to be effective as an index of a style of ideation—in George Klein's sense—based upon the analysis of such qualities in Rorschach verbalizations as flow, availability, and richness. "As a style RIRS should not necessarily measure 'repression,' but rather it should measure a property of ideation which predisposes one towards acts of repression, given the conditions for defensiveness."

They perused theoretical writings and examined a wide range of clinical and experimental data in the process of defining the construct of repression, in testing several nomological relations, and in validating the index against other relevant personality measures. Throughout the text they emphasize the limitations in their investigations and conceptualizations.

In the first chapter they review the relevant theoretical and ex-

perimental history of the concept of repression. They contend that verbal behavior is a basic mechanism of repression. Repression is defined as avoidant verbal behavior and as a limitation on ideational processes. This notion of repression is supported by such experimental studies as those on perceptual defense and selective recall in the face of unpleasant or threatening stimuli. Repression is seen also as having a relation to the stylistic cognitive characteristic of leveling-sharpening. They assert that labeling and categorizing of stimuli are primary functions of cognition. Much of the research reported in the book is directed toward clarifying the relationship between language and thought: "properties of language reflect properties of thought . . . . Repressive style is a consistent characteristic of an individual and it is manifested in vague, unelaborated language which is lacking in integration and flow."

Rorschach responses are viewed as rather uniquely useful in the derivation of an index of repressive style, for they are based upon a situation in which the test instructions and the test stimuli afford the subject minimal directions for labeling the stimuli. Therefore, a subject has no recourse but to respond in terms of his own frame of current and past experimential patterns to his introversive and extroversive orientations as represented by the test stimuli. They note that the evolution of Rorschach analysis has focused increasing attention on the language of the responses. The scoring system of the RIRS develops from the following general conception: "The more the verbalization of a Rorschach response reflects vague, impersonal and unelaborated thinking, and lacks integration and flow of ideas, the more repressive functioning has been manifest. The more the verbalization of the response is stated in specific, affectively toned terms and is characterized by a continued and developing flow of words, the less repressive functioning is indicated." Seven scoring categories were derived: specificity, elaboration, impulse responses, primary process thinking, self references, movement, and organization. Each free association is scored to indicate the lack of repression: thus, the lower the RIRS score, the more repressive functioning is said to be operative in the individual. The seven scoring principles are described in sufficient detail to enable the reader to use the index in his own research. An illustrative Rorschach performance is included, for which the scoring and tabulations are given on a specimen scoring sheet, and the index is calculated.

The chapter on reliability is a noteworthy contribution to the psychometric study of projective techniques. The authors report studies on scoring system reliability, including inter-scorer agreement, retest reliability in a broad range of test conditions and instructions, different sets of inkblots, and overvarying periods of time with a variety of subjects. They conclude that the RIRS is a

reasonably stable index and deals with a relatively enduring style of response.

In their explanation of the construct validity of the RIRS, they examined the relationship of the index to the Rorschach determinants of the Klopfer and Holtzman methods. A correlation was found between total quantity of verbalizations and the absence of repression. To test the question of whether verbosity is the underlying characteristic, the total number of words given by the same subjects to the Thematic Apperception Test (TAT) was compared with their RIRS; no relationship was found. RIRS scores were also compared to TAT Transcendence scores to shed further light on the psychological processes tapped by the Index.

RIRS has a low correlation with Stanford-Binet and Wechsler-Bellevue intelligence quotients as well as with need achievement measures for males from the TAT. Some sex differences in the correlation of RIRS with intelligence and achievement tests have been found. Developmental studies of the RIRS with data from the Gesell Institute and the Ledwith study of Pennsylvania students have suggested the value of evaluating the RIRS score in developmental terms to assess progression in differentiation of personality dimensions as well as subtlety and precision in the perception of self and others. Several studies are reported which compare RIRS scores with data from questionnaires which purport to measure anxiety. Clinical data and hypotheses suggesting a positive relationship between degree of repression and anxiety motivated the authors to study anxiety as a significant dimension in the construct validation of the term. Findings are equivocal, not only regarding correlations with anxiety indices, but also with MMPI indices of repression, sensitization, and self-ideal discrepancy. These findings point to the need for extensive testing of the sex differences in the functions reflected in the RIRS. One chapter is devoted to the discussion of available data on sex differences. Concurrent validity studies have produced either negative or inconclusive results.

To clarify the relationship between RIRS and subjects' inner versus outer frame of reference, the results of responses to field dependence, sensory isolation, and leveling-sharpening experimental situations were correlatied. Low but positive correlations were found between inner-frame-of-reference oriented individuals and high RIRS scores in contrast with environmentally oriented and low RIRS scores. The clinical validity of RIRS using differentially diagnosed textbook cases as well as other clinical material was found to be generally satisfactory in that RIRS scores were in the direction to be expected on the investigators' hypotheses for the various nosological categories examined.

The book which is well edited has an attractive format. It is well indexed.

In the reviewer's opinion this book merits serious reading by clinicians interested in testing techniques, by psychometrists interested in applying their discipline to the improvement of clinical techniques, and by students eager to learn about projective test reliability and validity methodology.

Cherry Ann Clark
The Meyers Clinic, Los Angeles

Interpersonal Competence and *Organizational Effectiveness* by Chris Argyris. Homewood, Illinois: Richard D. Irwin, Inc., 1962. Pp. 285.

The unintended, often dysfunctional, consequences of rational organization plans, rules, and procedures have received attention by many investigators, the latest being Argyris in this book. Similar in many general aspects, such as the self reinforcing aspects of such models, Argyris's study differs in the particular elements of the organization studied.

Specifically, Argyris agrues that since organizations are planned as rationalistic entities, emphasis is placed on the intellectual role that participants will play in organizations. In developing the intellectual role, organization planners tend to define rules, policies, and procedures which serve to guide people in their work. Neglected in this planning process is the rocognition that participants are capable of emotional and irrational behavior—behavior which is not legitimized by any aspect of the formal organization plan. No mechanisms are therefore specifically included in the organization structure which allow emotional behavior to come into the open and be dealt with as a paramenter. Implicit in such an approach to organization planning is a set of values which presumes that the only effective human relationships is best accomplished via rational behavior.

Argyris used this view of the organization plan to advance the following model. To the extent that participants accept and are guided in their behavior by the values implicit in the organization plan, they will tend to establish a social system with norms and defense mechanisms which prevent authentic interpersonal relationships and create a low degree of interpersonal competence. As interpersonal competence decreases in the organization, there will be a tendency toward conformity, mistrust, and dependence upon superiors within the managerial hierarchy which permeates the entire organization. In turn, there will be a decrease in effective decision-making, increase in organization defenses, and the creation of departmental centeredness and organizational rigidity—factors leading to a decrease in organizational effectiveness.

In an attempt to assess the validity of the model, interviews and

observations of the top executive system (18 executives) were made in the division home office of a large midwestern corporation. In addition to providing evidence to test the model, the interviews and observations were to serve as inputs into a diagnosis of the executive system. Rather than present any of the data uncovered in the study, suffice it to say that the data reported seem to support the model. That is, in all respects, the executives had developed a social system which reinforced rationalistic behavior and suppressed emotional behavior in interpersonal relationships, with a resulting low degree of interpersonal competence.

Armed with this preliminary confirmation, Argyris entered into a much more difficult project. After observing that interpersonal competence was low in the organization, Argyris used T-group training to see whether the executives' value system, and hence their behavior, could be changed toward higher interpersonal competence. After reviewing the evidence which consisted of observing the T-group sessions, questioning the executives, and observing their behavior, he concluded that the T-group experience had made the executives more aware of the importance of effective interpersonal relations and that there was a noticeable tendency toward higher interpersonal competence. However, because of limited time for observation and in light of inadequate instruments for measuring values, Argyris did not conclude that the T-group experience had definitely resulted in a change in the value system of the executives or that organization effectiveness had increased. The validity of these conclusions was reinforced by Roger Harrison's using independent measurements obtained from Kelly's Role Repertory instruments. However, it is difficult for the reader to reach the same set of conclusions as Argyris, since the major supporting evidence consists of dialogue from the T-group sessions and reports from interviews conducted after the sessions. One gets the impression that the excerpts included were specifically chosen to support the viewpoint of the author even though he makes a disclaimer on this possibility. Without the complete dialogue or, even more desirable, without being a witness to the sessions, it is difficult to reach any conclusion other than that given by the author.

As a contribution to the field of organization behavior, the book must be divided into two parts. The first part, the model formulation, presents many penetrating insights into the potential impact of a modern organization on its participants. In addition, many of the implications derived from the model provide testable hypotheses that can serve as the basis for additional research. Also provided is a highly informative discussion of T-group training as a mechanism for inducing change in organization behavior. The major limitation is the author's failure to provide operational definitions for most of the important concepts in his model. Particu-

larly noticeable is the vague definition of both interpersonal competence and organizational effectiveness.

The second part, consisting of the empirical work, is defective and incomplete in several respects. First, the data are represented in such a form that prevents a reader's independent judgement. Second, the data represent a very limited test of the model or of the ability to change an executive value system using T-group training, since just one portion of an executive hierarchy in one firm was used. No generality can be attached to the results. Third, the author did not have at his disposal any validated instruments by which to measure either the changes in executive values or the impact of value changes on executive behavior.

Argyris also noted a tendency among the executives involved in the T-group training to revert to their old, undesirable values with the passage of time. This "wearing off" tendency of supposedly internalized values leaves open an important question. Will the executives continue in their reversion to the old (pre-T-group) values or will the reversion stop at some point with a resulting net change in values? Argyris could not answer this question.

These empirical defects were all recognized by the author. Their citation in this review is not to serve as a criticism. Instead, it is to forewarn potential readers that the book does not provide either conclusive empirical evidence supporting the model developed or data on the success of the attempt to induce change in values of an executive system. Instead, it provides limited evidence, which when coupled with the model developed, will serve as inputs for further empirical testing.

D. J. Laughhunn
University of Illinois

Interviewing: Its Forms and Functions by Stephen A. Richardson, Barbara Snell Dohrenwend, and David Klein. New York: Basic Books, Inc., 1965. Pp.v + 380. $7.50.

The material contained in this book was derived from the Field Methods Training Program conducted at the Social Science Research Center of Cornell University. The book is divided into four parts: (1) The Interview as a Research Instrument; (2) Respondent Participation; (3) The Question-Answer Process; and (4) Interviewer and Respondent.

Part of the content stems from research done by the authors on the personality characteristics of sucessful interviewers, methods of selection and training, and the effects of various question types and their formulation on response quality and respondent participation. Studies by other researchers are also cited, but their methodology is not cirtically reviewed.

The interview for obtaining data and making judgements has been subjected to little research in contrast to the counseling interview, but a myraid of concepts and practices have accumulated. The few studies reported by the authors appear to contradict some of the prevalent assumptions and beliefs. For example, they present a study demonstrating that leading questions may enhance respondent cooperation and the validity of responses. They conclude that leading questions may, therefore, be useful and result in as accurate responses as nonleading questions. They further suggest conditions under which leading questions may be desirable. However, no empirical data are given to support their recommended conditions.

The authors caution that strong rapport is not always desirable. They give plausible examples for their recommendations but present no research data to support their contention. Similarly, they offer suggestions as to the appropriate visible characteristics an interviewer should have for various situations, but again supporting research data are lacking. This lack of supporting evidence is true for most of the recommendations offered by the authors for improving interviewing. This shortcoming is not really a condemnation of them as much as it reflects the general lack of empirical evidence for the folklore of interviewing practices. In comparison, competing books on interviewing techniques cite probably even less research evidence for the practices they advocate.

Interesting suggestions are offered for improving the determination of respondent samples, selecting interviewers, formulating questions, and obtaining respondent cooperation. However, the discursive style of writing makes it difficult to note readily these recommendations. The chapters not only seldom contain an introduction or brief preview of their content, but also do not end with a summary and set of conclusions. Consequently, it is not easy to read the book despite its relatively easy vocabulary and practical content.

Specific recommendations for interviewing procedures may be found more readily in other texts in the field such as Bingham, Moore, and Gustad's *How to Interview,* Fenalson's *Essentials in Interviewing,* or Kephart's *The Employment Interview in Industry.* Nor does this book present a theoretical framework as Kahn and Cannell do in the *Dynamics of Interviewing.* The patient and unharried reader doing field research in the behavioral sciences will find the book useful.

William E. Coleman
Coleman & Associates, Los Angeles

The Psychology of Learning and Techniques of Teaching by James M. Thyne. New York: Philosophical Library, 1963. Pp. 240.

Although based upon theoretical ideas in experimental psychology, this is a simply-written practical book on the use of learning theory in teaching. After building a learning model for analysis and discussion purposes, the author devotes roughly three quarters of the remainder of the book to how-to-do-it examples that apply to various teaching problems.

Thyne's working model for learning emphasizes four major conditions: cue, pilot cue, force, and tie. These terms refer (without specific mention of the fact) to concepts such as conditioned stimulus, unconditioned stimulus, drive, and reinforcement. It appears at first that the addition of these new terms instead of the use of conventional ones is unnecessary. However, since the author is very careful with his definitions, his terms probably avoid some negative transfer effects associated with disagreements about conventional terms.

Thyne's book is of value primarily to practicing teachers rather than to researchers. His direct reference to data is brief, and there is little to no discussion of controversy among important theoretical points. In some parts of the book there seems to be an emphasis on S-R contributions; in other parts, the emphasis is upon cognitive viewpoints. No deliberate attempt is apparent to connect examples to specific theories even though names such as Tolman, Guthrie, and Pavlov are used.

The major teaching problems discussed (with a chapter for each) are habit-training, habit-breaking, explaining, recall, and training in skills. As a rule, chapters begin with a definition of a problem (i.e., what are habits?). The discussion then tends to move carefully to related psychological concepts such as conditioning. From that point, an analysis is made of a classical experiment with implications for the learning model. With the establishment of the model, examples are mentioned with careful analyses as to how they tend to function.

Thyne is thorough in discussing tangential concepts such as motivation to aid the value of the learning model in teaching. Also, his attention to subpoints and their meaningfulness is reviewed with the insertion of summary statements and sections.

The last chapter of the book deals with transfer. The thesis is different from what is often provided in textbooks on educational psychology. Thyne, who does not desire to separate transfer from learning, mentions that the term *transfer* is now used because of historical precedent rather than because of psychological necessity.

The author has written a book that utilizes some of its own suggestions in the presentation of subject matter to the reader. It should be very useful in programs of teacher education, especially in

courses named "advanced educational psychology" in which a supplemental book on learning is valuable. The book lacks a discussion of the implications of research on meaningul verbal learning as proposed by Ausubel.

CHARLES M. GARVERICK
University of Illinois

The Preparation of Teachers, An Unstudied Problem in Education by Seymour B. Sarason, Kenneth S. Davidson, and Burton Blatt. New York: John Wiley & Sons, Inc., 1962. Pp. xv + 124. $3.95 ($1.95 paper cover).

As might be surmised from the title of this book, the authors consider that meaningful research has not been performed on the subject of teacher preparation. In defense of this title, the authors recognize that much discussion has been made, but the studies that have been involved with teacher preparation have been of a subjective nature both for hypotheses and for conclusions. It is asserted that empirical studies should be performed in which teacher education (the authors call it teacher training) is evaluated according to what teachers do and should be expected to do.

In the opinion of the authors, the teacher should be recognized as essentially an applied psychologist who is concerned with the learning process. The preparation of teachers, therefore, should be studied in ways that begin with a detailed analysis of the teacher-learner relationship. This concerns the understanding of teachers as observers, evaluators, and influencers of behavior.

A review is made of the controversy on teacher preparation that has existed between academic and professional groups. The authors use this review as an entry topic for elaborating the idea that subject matter knowledge does not insure effective teaching. This is not to say that knowledge is considered unimportant; it is simply that it alone is not sufficient. Neither, on the other hand, is a spirit of inquiry that may be an additional value in some courses emphasizing knowledge. Knowledge does not expose prospective teachers to problems of dealing with individual indifferences in students.

The remainder of the book describes classroom behavior and the preparation of prospective teachers to observe children. It is stressed that prospective teachers must have special preparation in classroom observation that is available as soon as possible after a decision is made to become a teacher. Such prospective teachers need help in perceiving their role as teachers. A tentative statement is made to the effect that the best teachers are able to perceive accurately greater differences among children. With this perceptiveness, they are better able to cope with children's needs.

In summarizing the book, one finds the major contribution to be the discussion of the program for handling classroom observation. The authors insist upon carefully selected introductory remarks before observations are made. They stress the importance of problems connected with an observer's failure to distinguish between overt behavior and an inference about covert behavior. Moreover, observers are not told what to look for, and judgments have to be justified in ways other than by the use of instructor-fed information.

In conclusion, the authors of this book are very critical of the subjective practices used in studying teacher preparation. It is thought that teacher preparation will remain uninvestigated unless studies are made of a teacher as a psychological diagnostician and tactician. However, the authors themselves have not presented a study with carefully-analyzed empirical evidence. Their position on teacher preparation, which is very interesting, possesses many recommendations that are thought-provoking. Again, however, many aspects of these recommendations represent overgeneralized conclusions. If documented with appropriate evidence, the recommendations could become a much more important contribution to education.

Many college teachers involved with teacher preparation should be aware of the ideas expressed in this book. It is doubtful that the book will be used as a text by undergraduates in educational psychology or other courses in teacher preparation.

CHARLES M. GARVERICK
University of Illinois

Concepts of Development in Early Childhool Education by Peter Neubauer (Editor). Springfield, Illinois: Charles C. Thomas, 1965. Pp. ix + 149.

This book is an exciting treatment of the most crucial and recent concerns in early childhood education; Neubauer, who is director of the Child Development Center, which has a nursery school and a clinic for the treatment of children of preschool age, reports the proceedings of a two-day Institute given at the Center in September, 1964.

The major purpose of the Institute was to clarify ideas related to child development, subsequently the bases for criteria in the assessment of development. The editor states that the wide gap between psychiatry and, what we hope education will do in the socialization processes for children, needs to be narrowed. His concern is that many children are going by, unnoticed by the teacher, when the child actually may need help. "Without closer coordination, we may be running the risk of either too much treatment or not enough. The key to such coordination is the concept of development

on which we have not yet arrived at wide areas of agreement" (p. vii).

Four papers were presented: (1) "Stages of Intellectual Development and Their Implications for Early Childhood," by David Ausubel M. D. and Ph.D., who represented a medical and educational psychological point of view; (2) "Language and Development," by Lili Peller, who is a psychoanalyst; (3) "Developmental Considerations of the Nursery School Experience," by Irving Sigel, who is a research director of Merrill-Palmer Institute; and (4) "Death Fear and Climax in Nursery School Play," by Jules Henry, who is an anthropologist and sociologist.

Stages of intuitive functioning and logical reasoning presented by Ausubel was an astute work; it gave much substance to a rigorous discourse by a reacting panel which consisted of people highly knowledgeable in their fields. Ausubel included four types of intuitive processes which he feels are often ignored by many; if those four types were examined as separate phenomena semantic and pedagogical clarity might result. He also discussed the similarities and differences between adult's and children's thinking.

Biber takes issue with what she feels is an assumption on the part of many who seem to feel that it is the job of educators to deal with an earlier mean emergence of cognitive functioning of young children. Extrapolating the non-cognitive from the cognitive development of children in their early years is one of the dangers of contemporary educators. A kind of equilibrium needs to be developed for the normal functioning of young children; imbalance of cognitive development can lead to psychological disorders.

Peller, in her presentation on language development, indicated the circular character of early play: "A motion or a sound produced by the infant becomes the stimulus for a new act on his part. Because the child has played with himself and answered his self-created stimulus, he is more ready to play back and forth with his mother. On the other hand, a well-cared for infant is more persistent in his self-play, because the *mother* has played with him" (p. 68). The human face is very important in a child's learning language; the child responds with great affect to the mother's face.

She also indicates that there is an important area of "contacts with the environment, which the child himself initiates, and the even huger field of self-stimulation," (p. 73) which many people seem to forget when they discuss early childhood education; the tendency is to think either of maturation processes or of instruction in relation to children's learning and development. "Spontaneous exploration and circular stimulation may even account for most of the learning of young children" (p. 73).

Sigel's paper discusses the child's separation from home to

school, the physical world of the nursery school, and its contribution to the intellectual aspect of children's development. Very little research has been done on the nursery school examined from the points of view of his framework. He does emphasize, however, that the nursery school teacher not be considered a parent surrogate, that children are reality-oriented enough to know she is *not* a substitute parent, that she is a helpful adult. Educators need to be aware of this concept, too.

Children are socialized by social rewards which are given for "impulse control, respect for others, achievement striving, and reality orientation. In other words, in the nursery school, social rewards contribute to the child's acquisition of role behaviors" (p. 92).

Experimentation with roles occurs in the "nursery school experience in two ways: (1) the real-life requirements necessitated by the child's ship from a homebound individual to the school child; and (2) dramatic play, either spontaneous, or as part of the teaching program" (p. 93).

Henry suggests that underlying anxieties about annihilation reveal themselves in children's enactment of commercialized television's "ceremonialization of death, fear, hostility," (p. 123) which "provides a spectrum of mass-produced roles which children of nursery school age readily make part of their play, thereby making possible a certain kind of special interchange among themselves" (p. 123). He feels that since fear is a part of the culture, the notion has to be dealt with: "All we can expect to do is to enable other types of human potentialities to emerge in the process of acting out fear. Impulses to mastery, to protectiveness, to love, and to clear reasoning readily emerge in the danger situation, and ought to be fostered" (p. 123).

The tendency to project adult notions into interpretation of children's behavior, overt and covert, has been prevalent in research. Children's notions about death, sex, and mastery, are mainly translated through the adult perspective, but few research studies are available to give us satisfactory data on these important questions.

Even though consensus was not reached, congruent to the major purposes of the Institute, the book itself, with its discussions, reactions, and suggestions to the major papers presented is an excellent source, rich with ideas, and valid in its most recent trends of contemporary and crucial concerns in early childhood. This reviewer recommends heartily that anyone interested in the field see this book; not only is it fine reading, but it is also a rich source for further research.

EDYTHE MARGOLIN
University of California, Los Angeles

Psychodrama Explained by Samuel Kahn. New York: Philosophical Library, Inc., 1964. Pp. vi + 77.

This slim volume, unencumbered by reference material, explains the nature and purpose of psychodrama. The text is preceded by an introduction by J. L. Moreno. The psychodrama is viewed as a "form of psychotherapy in which the participants enact or re-enact situations that are of emotional significance to them." Its use is not restricted just to individuals who are in an abnormal state. Kahn believes that almost anyone can gain insight into his problems through this procedure. The psychodrama employs verbal communication but also includes various non-verbal means of communicating such as gestures, movement, and dancing.

Kahn examines several sources that have contributed to the development of this procedure. Among these are Freud's "social instinct" and Trotter's idea that a group is a "biological extension of all higher organisms." He also traces Moreno's contribution to the psychodrama. The general contrasting of psychodrama with other procedures enables the reader to see how it relates to the total therapeutic structure. According to Kahn, one of the chief values of psychodrama is the spontaneity which it generates within the individual and the group. This spontaneity frees forces that enhance the therapeutic process in all who are witness to the scene.

The book has six chapters or subdivisions. The first two chapters provide background material. The third chapter, "The Psychosomatic Picture", examines the need for medical and psychological histories of the participants in the drama. It is based on the premise that we all are neurotic to a certain degree. The fourth chapter is more than a glossary of terms. It is the examination in some detail of the ideas, phrases, words, characters, concepts, and techniques that are used by those who employ psychodrama as a therapeutic procedure. The fifth chapter provides a meaningful synthesis of the concepts and techniques of the total process. The last chapter gives an example of a psychodrama session in action.

To a certain degree the book tends to appear "simple" in nature. However, it does demand a degree of psychological sophistication to read it with meaning. For example, the psychodrama is not another role playing technique but a form of depth therapy. This subtlety takes a little time to be accepted by the general reader. Although Kahn has taken care in "explaining" psychodrama, only time will tell whether the procedure will have universal usage.

HENRY KACZKOWSKI
University of Illinois

Coming of Age in America by Edgar Z. Friedenberg. New York: Random House, 1965. Pp. xii + 300. $5.95.

Among the frequent criticisms of educational research is that the

results are more often found on a library shelf under a layer of dust than in action. The usual nostrum begins with the need to rewrite the research report in the vernacular and to project the results beyond the safe hedging of fiducial limits into specific programs for action. Dr. Friedenberg has apparently attempted to meet the prescription. It will be interesting to observe the results.

A total of 225 students from nine high schools was interviewed to secure their opinions on reasonable solutions to fictional situations involving "typical" high school students interacting with the organization and the staff of a mythical school. From their responses, some of which were secured through a modified Q-sort and some through a hermeneutic analysis of the interviews, Friedenberg discusses the formation and operation of adolescent values in the American high school.

The initial analysis presents the adolescent as a native under colonial government. Programs for the adolescent are not initiated for the sake of the adolescent but rather to get him to abandon his barbarianism and to acquiesce to the patterns of adult life. By definition, the adolescent is inferior. Moreover, in every task assigned to him by the school, whether academic, social, or organizational, his inferiority is emphasized. Assimilation into adult life is achieved at the cost of individual freedom, human dignity, and elemental justice.

The theoretical structure of the research is never clearly presented. In fact, the author himself points out that he is supplementing the research with other experiences and conclusions and that he is using some of his research data to illustrate his points. The definition and dynamics of values as used in the book are never specified. Nor do they become apparent from the continuity. Opinions and attitudes are often used as synonyms for values. Furthermore, the relations to overt behavior or action are not clear. However, the book is a popular rather than a scientific report; one should not complain about the absence of modern plumbing when one intends to live with the natives in the bush.

Following the presentation of the study, which at times seems excessively long on data and short on interpretation, Friedenberg summarizes his personal views on the treatment of the adolescent. He strongly disapproves of compulsory school attendance; the inability of adults to design and to operate a workable social order is no reason to confine the adolescent until he accepts mediocrity. The profession of education must begin to base its authority upon its own competence to assist the adolescent in his development rather than to rely upon tradition, physical force, Machiavellian coercion concealed beneath labels of "psychological" or "guidance" activities. The school must aspire to the task of education, which has a goal of helping the individual to find meaning in his own life and,

to become more sensitive to the meaning of the lives of others, and to desist from the dowdy, prosaic, essentially goaless task of providing inadequate custodial care for the natives until they are tired of being restless.

The book is an interesting one; it should generate a modest amount of controversy. It does not present a pleasing picture of the American high school or of its students and its staff. Unfortunately, all three of these educational segments have been criticized so often that, like the natives under colonial government, they accept their lot with little complaint and are advised by the master to be grateful for the few crumbs lest these also be taken from them.

Gerald T. Kowitz
University of Houston

Readings for Educational Psychology by Ellis Batten Page (Editor). New York and Burlingame: Harcourt, Brace & World, Inc., 1964. Pp. xii + 404.

This book of readings was prepared by Page in consultation with Lee J. Cronbach. As such, the readings are especially designed to accompany Cronbach's *Educational Psychology,* Second Edition (Harcourt, Brace & World, 1963). Nevertheless, some of the selections are valuable readings appearing after Cronbach's book. Others are taken from sources not appropriate for listing in his text, as they are relatively unavailable to students.

The most logical question to be considered is, "Does the collection of readings adequately sample the field?" This question particularly applies to educational psychology, since that discipline reflects a variety of professional interests, rules of evidence, and methods of approach. Page has deliberately chosen some great thinkers from many fields—psychology, anthropology, sociology, and philosophy, all of whom are, of course, "educators!" The authors include Paul Mussen, Else Frenkel-Brunswick, Jackson and Getzels, Havighurst and Neugarten, Jerome Bruner, and B. F. Skinner. Whereas no book of readings could be expected to include all areas of the field, Page has presented the grand issues of the field quite adequately.

Another appropriate question would be concerned with how has the editor combined and arranged the book for maximum clarity and meaningfulness. Page lists eighteen subjects with two readings per subject for a total of thirty-six readings. Titles of subjects include, "How Psychology Contributes to Education," "Communicating Knowledge," "Purposes and Aspirations," "Judging Performance," and the like. Apparently Page has put more emphasis upon selecting a good reading list and less emphasis upon selecting several readings on a particular topic. However, the loss in depth of any particular subject is more than made up by the selection of worthwhile articles.

As might be expected in a readings designed to accompany Cronbach's book, the over-all emphasis is toward rigor; the editor takes the view that educational psychology is at its best an empirical social science. The readings chosen are articles involving manipulation of variables and quantification of results. If education is to be defined as an art and psychology as a science, then it appears that this book of readings leans more toward psychology than it does toward education. In spite of the "solid" approach to the field, some "unscientific" readings of note are included.

Finally, in the introduction of each subject area, no prefatory statement is made by the author; instead a brief headnote describing the reading, raising some questions, and making some suggestions appears at the beginning of the article proper. This volume is certainly a refreshing change from some books in education which one reads and then wonders what one has read. The editor has successfully chosen and arranged the selections such that the grand issues in educational psychology are cogently stated and the material is highly factual; yet the readings are not so trivial or overwhelmingly statistical that the student becomes dismayed.

PHILIP S. VERY
University of Rhode Island

ERRATUM

An error has been discovered in the list of references for the article "Acquiescence in the MMPI?" by Leonard G. Rorer and Lewis R. Goldberg which appeared in the autumn 1965 issue. On page 817 the reference to the article by L. G. Rorer entitled "The Great Response Style Myth" lists the wrong journal title. The article actually appeared in the *Psychological Bulletin* instead of the listed *Psychological Review*.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

SINGLE AND MULTIPLE HIERARCHICAL CLASSIFICATION BY RECIPROCAL PAIRS AND RANK ORDER TYPES

LOUIS L. McQUITTY
Michigan State University

Hierarchical Classification by Reciprocal Pairs is a statistical method for classifying people (institutions or other objects) into a hierarchical system based on indices of association between them (McQuitty, 1964).

The method selects initial pairs of objects which are indicative of types in terms of common characteristics and generally expands and improves the types by incorporating additional objects and by eliminating characteristics which are irrelevant to the descriptions of the types. The method realizes both types of any size and successive levels of classification while at the same time limiting itself to reciprocal pairs.

A pair of objects i and j are reciprocal if i is most like j and j is also most like i $(i \rightleftharpoons j)$; but not if only i is most like j, or if only j is most like i; i does not need to be most like j just because j is most like i and vice versa.

Reciprocal pairs have two advantages: (a) the very fact that the members are reciprocal is generally some evidence that the classification is valid; and (b) every matrix of interassociation between objects has at least one reciprocal pair; the highest index in the matrix necessarily mediates between reciprocal objects, and the matrix may have several or many reciprocal pairs of objects. If Index ij is the highest index in the matrix, then i is highest with j and j is also highest with i. In this latter case, however, the reciprocity may be due exclusively to the fact that the index is the highest entry in the matrix.

If the successive matrices each yield only one reciprocal pair, the resulting classification could be due to chance alone; additional evidence of validity is required, such as by a repeated and independent study. If, however, the successive matrices produce several reciprocal pairs, this, in and of itself, is some evidence of validity, but nevertheless not sufficient for many situations.

No matter how few or how many reciprocal pairs are found in a matrix, the members of each pair are classified together and in the next step of the analysis each pair is treated as a single unit in the classification.

A new matrix is formed. It is composed of interassociations between: (1) objects with objects, (2) object pairs with objects, and (3) sometimes object pairs with object pairs if the original matrix contained more than one reciprocal pair.

If pairs of objects with pairs of objects or a single object with a pair of objects yields a reciprocal pair, then in the next matrix this reciprocal pair is treated as a single unit. In the general case each pair is treated as a single unit in the next matrix no matter how many objects have been involved in building it up from the analysis of previous matrices.

Of the three kinds of interassociations mentioned above, those of Item (1) are available from the original matrix. Those of Items (2) and (3) can be computed using the *classification assumption* (McQuitty, 1960), the *similarity index* (McQuitty, 1955a), or the *corrected agreement score* (McQuitty, 1956).

The *classification assumption* assumes that the interassociation between two categories is equal to the smallest index when every object from among the two categories is paired with every other one and their indices computed.

The *similarity index* uses the indices for all pairs obtained when one object of each pair must come from each category; it is the average of the indices for these pairs.

The *corrected agreement score* is designed to correct the agreement score, n_{ij}, for chance.

Hierarchical Classification by Reciprocal Pairs Using the "Pure-Type" Assumption

The Assumption

A difficulty with the above method is the complication involved

in computing indices of association between pairs or between an object and a pair of objects.

We have developed a new and simpler version of Hierarchical Classification by Reciprocal Pairs. It is based on several assumptions: (1) Every object is an "imperfect" type as distinct from a "pure" type; only "imperfect" types exist in reality, and "pure" types exist only in theory. (2) There are fewer "pure" types than "imperfect" types; each "pure" type is represented in reality by two or more "imperfect" types. (3) The characteristics of "pure" types are approached but never quite realized by classifying "imperfect" types into internally consistent categories, and determining their common characteristics; the validity of representation of a "pure" type increases as the number of "imperfect" types representing it increases. (4) "Hierarchical" types include all of the types realized in classifying "imperfect" types into larger and larger, internally consistent categories; they are the types intermediate between those of reality and theory, "imperfect" and "pure". (5) Deficiencies in the validities of "imperfect" and "hierarchical" types as representatives of "pure" types are insufficient to cause erroneous classification when the method of classification is based on the isolation of internally consistent categories, as in the case of Hierarchical Analysis by Reciprocal Pairs. It is unnecessary to correct for chance errors in the indices between "imperfect" types as we classify them into larger and larger, internally consistent categories in approaching "pure" types.

The steps in Hierarchical Analysis by Reciprocal Pairs, using the "Pure-Type" Assumption, are as follows:

(1) Compute an index of association between every "imperfect" type with every other "imperfect" type to yield a matrix of interassociations between "imperfect" types.

(2) Isolate the reciprocal pairs.

(3) Specify the common characteristics for each reciprocal pair. These define a "hierarchical" type. It is now possible to compute an index of association between either, (a) two "hierarchical" types, or (b) between a "hierarchical" type and an "imperfect" type.

(4) Substitute the "hierarchical" types for the "imperfect" types which produced them.

(5) Prepare a new matrix of interassociations between types,

using only the retained "imperfect" types and the "hierarchical" types which replaced the discarded "imperfect" types.

(6) Isolate reciprocal pairs.

(7) Continue these steps until the analysis is completed.

An Illustration

The method is illustrated in analyzing Table 1, which shows

TABLE 1

Agreement Scores between Companies

	A	B	C	D	E	F	G	H
A		29	16	16	14	6	11	7
B	29		17	17	13	6	8	10
C	16	17		26	10	8	9	13
D	16	17	26		10	12	11	11
E	14	13	10	10		21	17	13
F	6	6	8	12	21		19	17
G	11	8	9	11	17	19		24
H	7	10	13	11	13	17	24	

Note—Data from McQuitty, 1954

agreement scores between eight companies in terms of their union-management relations as assessed in terms of 32 dichotomized variables. Two companies agree on a variable if they are either both above or below the median, but not if one is above and the other is below.

The entries for the reciprocal pairs are underlined; they are AB, CD, EF, GH; two construction companies, two trucking companies, a grain processing and a metal products company, and two garment manufacturing companies, respectively.

As the entry between A and B shows, these two companies have common answers on 29 of the 32 variables; they disagree on three variables.

In the next step of the analysis, each "hierarchical" type, such as AB, is assessed on the items on which its two members agree, exclusively. This means that Types AB, CD, EF, and GH were assessed on 29, 26, 21, and 24 items respectively in the next stage.

The next step was to prepare a matrix (Table 2) reporting the agreements between Hierarchical Types AB, CD, EF, and GH. In order for two types such as AB and CD to agree on an item both types must be assessed on that item and all companies of the two types would have to have the same answer on the item, all either

TABLE 2

Agreement Scores between Hierarchical Types

	AB	CD	EF	GH
AB		13	4	5
CD	13		4	6
EF	4	4		10
GH	5	6	10	

above or below the median but not some above and others below.

The entries for the reciprocal pairs, underlined, show that Hierarchical Types AB and CD join to form a higher level Hierarchical Type ABCD, and likewise EF and GH to form EFGH, to yield the results shown in Figure 1.

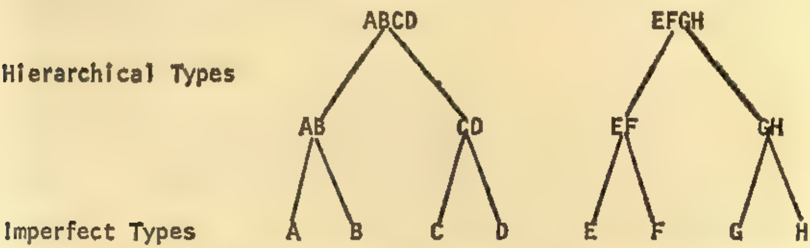


Figure 1. Types of Companies in Terms of Some Union-Management Characteristics

It is a characteristic of the data and not the method that the classification is by twos. The method requires only that the reciprocal pairs of the first matrix be by twos. Other "hierarchical" types can contain any number of members from three up; a "hierarchical" type of two members from the first matrix, for example, could join a single "imperfect" type of the second matrix.

An Interpretation

The example illustrates how the method conforms to its design and improves on the validity of the types as it proceeds. This fact is illustrated with respect to Type EFGH. In Rank Order Typal Analysis, where a test of internal consistency was applied, this category of four companies failed to qualify as a type. (McQuitty, 1963). In the present study, on the other hand, it did qualify even though a test of internal consistency was applied. It did so because each Type EF and GH was improved as a representative of Type EFGH, and the test of internal consistency was applied to them

rather than to the individual companies (the "imperfect" types) as in Rank Order Typal Analysis. The Types EF and GH were improved by eliminating from their respective descriptions items on which their two members did not agree. The test of internal consistency was satisfied in the sense that the two types EF and GH produced a reciprocal pair, which is a less stringent criterion than used by Rank Order Typal Analysis.

The Principle of Maximum Classification

Additional assumptions can be used to (a) increase the qualifications of "hierarchical" types in order to qualify as "pure" types, and (b) assist in developing multiple "pure" type methods.

"Pure" types are assumed to be most adequately approached when the number of objects classified into a category times the number of characteristics on which they agree is maximized. If this product is as large (or larger) for a proposed category than for any "hierarchical" types entering the proposed category, then the proposed category is retained; the analysis proceeds using the proposed category as a "hierarchical" type.

Otherwise the proposed category is rejected; the "hierarchical" types which were to enter it remain and continue in the analysis as two separate "hierarchical" types.

In order for the two separate "hierarchical" types to classify elsewhere in the subsequent analysis, each of them must be allowed to qualify as reciprocal with some other type. This is done by ignoring for the balance of the analysis the high index which exists between them. Either of them, i or j, is said to be reciprocal with some other type, say k, if either $i \rightleftharpoons k$ or $j \rightleftharpoons k$, except for ij.

The logic of the above approach is this: some reciprocal pairs probably occur by chance alone. Consequently, another check is sometimes needed to determine whether or not reciprocal pairs do in fact reflect "pure" types. Consequently, they are assumed to form "pure" types, if, and only if, in addition to being reciprocal, they satisfy the *principle of maximum classification* as outlined above.

The *principle of maximum classification* is not generally applied when either two "imperfect" types or an "imperfect" type and a "hierarchical" type yield the proposed category; it is assumed that an "imperfect" type might win over a proposed category due primarily to excessive, irrelevant characteristics. Once "imperfect"

types have joined to form a "hierarchical" type the number of irrelevant characteristics has usually been reduced, and generally to a considerable extent if there was originally an unusually large number of them; erroneous decisions by the principle are therefore no longer anticipated.

Multiple Hierarchical Analysis by Reciprocal Pairs for the Isolation of Independent Types

The *principle of maximum classification* can be used to convert Elementary Hierarchical Classification by Reciprocal Pairs into a multiple classificatory method.

Assume that the principle has been applied as outlined above and that it was not satisfied, then the proposed category is rejected. Furthermore, the two "hierarchical" types which were to form it become "finalized." The common characteristics of each such "finalized" category are assumed to be the best possible, even though somewhat imperfect representatives of a "pure" type.

Both the finalized categories and the objects which compose them are continued in the analysis, but they are first separated completely in terms of the characteristics which describe them. Each finalized category is defined exclusively in terms of the common characteristics of the objects which compose it; each of these objects, on the other hand, is defined exclusively in terms of the characteristics not common to the objects which make up the category.

As a consequence each reduced object has a zero agreement with the finalized category in which the object was originally classified.

As a consequence of this procedure, both finalized categories and the reduced objects can classify in a later analysis.

The above *principle of maximum classification* was developed as a component part of the development of Multiple Classification by Agreement Analysis (McQuitty, 1955b, 1957, 1962; Hemingway, 1961). This is but one principle, others might be developed and shown to be better for specified purposes.

An Illustration

Multiple Hierarchical Classification by Reciprocal Pairs for the Isolation of Independent Types can be illustrated with the data reported in Table 1. In fact, with this data, the procedures of this method and the earlier method are identical for the analysis of

Tables 1 and 2; the unique features of the current method are introduced only after Table 2 has been analyzed.

The *principle of maximum classification* is applied to the two "hierarchical" types yielded by Table 2, Types ABCD and EFGH.

The members of Type ABCD agree on 13 items. (Table 2) and there are four members. Consequently, the classification capacity is $13 \times 4 = 52$. The classification capacity of one of the two types out of which they developed is larger than this number. The members of Type AB agree on 29 items (Table 1) and this two-member type therefore has a classification capacity of $29 \times 2 = 58$. Analogously, the classification capacity of Type CD is $26 \times 2 = 52$. In terms of the *principle of maximum classification,* Type ABCD is rejected in relation to Type AB.

Analogously, Type EFGH is rejected in relation to each Type EF and GH.

The finalized categories of the first analysis are AB, CD, EF, and GH. The description of each A and B was reduced by their 29 common characteristics leaving the category, AB, described in terms of these 29 items and each reduced object, A′ and B′, described in terms of the three remaining items (on which they disagree). The other three categories were treated in the same fashion. Using this revised data the agreement scores for the matrix of Table 3 were determined.

As Table 3 shows, there is little left here on which to have additional classifications; the first analysis was relatively exhaustive for the data of this study.

The Table does illustrate that the method would continue to work even with this data. All of the proposed classifications between any two types, such as CD and GH would be rejected by the *principle of maximum classification.* This would continue until the agreement scores of three in the lower right one-quarter of the matrix, mediating between reduced objects, were the highest entries in the matrix. They would then yield the additional classifications: A′G′, B′H′, C′E′, D′F′, D′G′, E′G′, and F′H′. Even these classifications (based on few characteristics) hold a consistency with the first classification. In the first classification A and B formed a type as did also G and H, and here A′ joins G′ and B′ joins H′. This same type of relationship exists in all of the new classifications except for D′G′.

TABLE 3

Agreement Scores between Finalized Categories and Reduced Objects

	AB	CD	EF	GH	A′	B′	C′	D′	E′	F′	G′	H′
AB		*	4	5	0	0	2	2	8	1	5	2
CD	*		4	6	0	1	0	0	3	3	1	3
EF	4	4		*	1	0	0	2	0	0	3	1
GH	5	6	*		0	0	1	1	1	4	0	0
A′	0	0	1	0		0	1	1	1	1	3	0
B′	0	1	0	0	0		1	1	1	1	0	3
C′	2	0	0	1	1	1		0	3	1	1	2
D′	2	0	2	1	1	1	0		1	3	3	1
E′	8	3	0	1	1	1	3	1		0	3	2
F′	1	3	0	4	1	1	1	3	0		2	3
G′	5	1	3	0	3	0	1	3	3	2		0
H′	2	3	1	0	0	3	2	1	2	3	0	

* Rejected earlier by principle of maximum classification

Other Possible Outcomes

The above outcome need not have ensued with a larger set of data. Suppose we had had the set of data shown in Table 4, which incorporates Table 2. Types ABCD and EFGH would be rejected just as reported above.

TABLE 4

An Expansion of Table 2 to Include a Hypothetical Type IJK

	AB	CD	EF	GH	IJK
AB		13	4	5	12
CD	13		4	6	9
EF	4	4		10	8
GH	5	6	10		7
IJK	12	9	8	7	

With Type ABCD rejected, Type ABIJK becomes a candidate, and it qualifies in relation to Type AB; it has a classification capacity of $12 \times 5 = 60$, as compared with the capacity of 58 for AB, as reported above. Assume that the classification capacity for Type IJK is 42 ($3 \times 14 = 42$). Category ABIJK would have been accepted as a type.

Multiple Hierarchical Analysis by Reciprocal Pairs for the Isolation of Intersecting Types

Hierarchical Classification by Reciprocal Pairs always yields a typology, a hierarchical system of types of objects, with each type

defined in terms of the characteristics which its objects have in common. In addition to this predominant typology, there may be other typologies which are obscured by the dominant one, but which nevertheless can be isolated by appropriate procedures.

An Illustration

Consider, for example, the typology shown in Figure 1. The characteristics common to any one of the "hierarchical" types, say CD (two trucking companies) might prove to express itself in the "imperfect" types and there yield a typology somewhat different from the predominant pattern. This possibility can be examined as follows.

Determine the items on which Companies C and D, the members of Type CD, agree, and using only these items compute a matrix of agreement scores for all companies as shown in Table 5. Since Companies C and D agree on only 26 of the 32 items, Table 5 is

TABLE 5

Agreement Scores between Companies Using Only the Items on Which C and D Agree

	A	B	C	D	E	F	G	H
A		25	13	13	10	5	8	6
B	25		14	14	9	4	7	7
C	13	14		26	7	7	7	9
D	13	14	26		7	7	7	9
E	10	9	7	7		19	14	12
F	5	4	7	7	19		14	14
G	8	7	7	7	14	14		22
H	6	7	9	9	12	14	22	

based on only these 26 items and the highest possible entry in the matrix is 26.

The data of Table 5 were analyzed by Hierarchical Classification by Reciprocal Pairs, using the "Pure-Type" Assumption. The analysis yielded the same types as the analysis of Table 1 except for one important difference. Both analyses yielded Types AB, CD, EF, GH, ABCD, and EFGH, as shown in Figure 1. However, in the case of Table 1, both ABCD and EFGH, failed to qualify as types in terms of the *principle of maximum classification.* In the case of Table 5, Type ABCD qualifies and EFGH continues to disqualify, i.e., in terms of the additional requirement of the *principle of maximum classification.*

In the last analysis, Types AB, CD, and ABCD agree on 25, 26, and 13 items respectively, yielding classification capacities of 50, 52, and 52 respectively, and thus barely qualifying Type ABCD.

Types EF, GH, and EFGH agree on 19, 22, and 10 items respectively, yielding classification capacities of 38, 44, and 40 respectively; Type EFGH fails to qualify in terms of the *principle of maximum classification.*

The items of all other pair-wise types were each utilized in an identical fashion as just reported for those of Type CD. They all produced the same structure as shown in Figure 1, except the items of Type AB; they conformed except that they failed to yield Type EFGH. In none of these other three cases did either Type ABCD or EFGH qualify as a type in terms of the *principle of maximum classification;* only the items of Type CD (trucking) resulted in Type ABCD qualifying in terms of both internal consistency and the *principle of maximum classification.*

The items of Type CD, EF, and GH each produced Type EFGH with a classification capacity of 40, but always with a failure to qualify in terms of the *principle of maximum classification.*

An Interpretation

The typology reflected by Type CD (trucking) is more representative of Type ABCD than is the typology reflected by Types AB (construction), EF (grain processing and metal manufacturing), and GH (garments), and is equally as effective as these other pair-wise types in producing Type EFGH; of the four pair-wise types, Type CD contains the core most representative of the entire typology as is outlined in Figure 1.

Hierarchical Classification by Rank Order Types

Hierarchical Classification by Rank Order Types (McQuitty, 1963 and 1964) is an elaboration of Hierarchical Classification by Reciprocal Pairs (McQuitty, 1964). Reciprocal pairs are internally consistent categories of only two objects; each object of every category is more like the other object in the cateogry than it is like any object of any other category.

Internally consistent categories can, however, be composed of more than two objects. In this case, every object of every category is more like every other object of that category than it is like any

object of any other category. Hierarchical Classification by Rank Order Typal Analysis isolates all of the internally consistent categories of each matrix, irrespective of the size of the categories.

There is but one exception to this latter principle. The method will not accept Category X of x objects as satisfying the definition of internal consistency if X includes any $x - y$ objects (y less than x by at least 2 or more) which do not satisfy the above definition. In other words, the categories must not only be internally consistent to qualify as types, but in addition all of the sub-categories which were considered in building them must also qualify. If this were not true, the method could prove to be extremely laborious with some sets of data.

A Comparison of Analysis by Reciprocal Pairs and Rank Order Types

An analysis by Rank Order Types is not necessarily more rapid than by Reciprocal Pairs just because the former method can often achieve a solution by the analysis of fewer matrices; each analysis of a matrix by that method is usually much more elaborate than by Reciprocal Pairs. In fact, the latter method is generally less laborious, except for highly-structured data.

Another advantage of the latter method is that it is presumed to improve "imperfect" types and "hierarchical" types very gradually as it proceeds. The former method attempts to improve them in terms of larger units and can in fact take such large steps that some "imperfect" types and even some "hierarchical" types might be left out of possible improvements. The less well-structured the data the more apt this is to occur.

Analysis by Rank Order Types is preferable for highly-structured data and by Reciprocal Pairs for less well-structured data.

Summary

This paper outlines definitions of "pure," "imperfect," and "hierarchical" types and uses them to improve Hierarchical Classification by Reciprocal Pairs and expand the method to isolate multiple typologies of both *independent* and *intersecting* types.

REFERENCES

Hemingway, Peter Wing. Multiple Agreement Analysis. (Ph.D. thesis, Michigan State University Library, 1961.

McQuitty, L. L. Pattern-Analysis: A Statistical Method for the Study of Types. In W. E. Chalmers, M. K. Chandler, L. L. McQuitty, R. Stagner, D. E. Wray, and M. Derber (Eds.) *Labor Management Relations in Illini City,* Volume II. Champaign, Illinois: Institute of Labor and Industrial Relations, University of Illinois, 1954.

McQuitty, L. L. A Method of Pattern Analysis for Isolating Typological and Dimensional Constructs. University of Illinois, Contract No. AF 33(038)–25726, December, 1955a, *AFPTRC Research Report TN–55–62.*

McQuitty, L. L. Multiple Classification by Agreement Analysis in the Isolation of Types of Job-Knowledge Item-Responses. University of Illinois, Contract No. AF 33(038)–25726, November, 1955b, *Memorandum Report A–29.*

McQuitty, L. L. Agreement Analysis: Classifying Persons by Predominant Patterns of Response. *British Journal of Statistical Psychology,* 1956, 9, 5–16.

McQuitty, L. L. Isolating Predictor Patterns Associated With Major Criterion Patterns. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1957, 17, 3–42.

McQuitty, L. L. Hierarchical Syndrome Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 293–304.

McQuitty, L. L. Multiple Hierarchical Classification of Institutions and Persons With Reference to Union-Management Relations and Psychological Well-Being. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 513–531.

McQuitty, L. L. Rank Order Typal Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 55–61.

McQuitty, L. L. Capabilities and Improvements of Linkage Analysis as a Clustering Method. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 441–456.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

COMPARISON OF TWO APPROACHES—JAN AND PROF—FOR CAPTURING RATER STRATEGIES[1]

ROBERT J. WHERRY, Sr. and JAMES C. NAYLOR
The Ohio State University

In an earlier paper the present authors (Naylor and Wherry, 1965) have described the use of the Bottenberg-Christal JAN technique (Bottenberg and Christal, 1961) for the isolation of rater policies. In that paper we reported in detail the JAN outcome for one of four air force specialties analyzed. Actually four specialties were studied and the present paper is based upon the complete study. The four specialties consisted of two supervisory levels of a mechanical specialty and two versions of an administrative (housekeeping) type of specialty, both versions being at the same supervisory level.

The JAN technique is based upon defining the capturing of rater policy as the extent to which one can predict the actions of a rater from the human characteristics of the data he is being required to evaluate. With this definition we have no disagreement, but the process of capturing policies and of identifying (describing) them and similarities and differences among them are two distinctly different problems, as we stated in the earlier paper.

The JAN technique may tell us that raters 1, 13, 22, and 47 use the same strategy, indicating that the same regression equation applied to the stimuli works almost as well as would four separate regression equations in predicting the rank order assigned to the ratees by these four raters. Raw score regression equations are, however, notoriously difficult to interpret, due to such internal in-

[1] This research was supported in part by the Air Force Systems Command, Contract AF41 (609)-1596. The opinions expressed in this paper are those of the authors and not necessarily those of the Air Force.

teractions between predictors and criterion as are reflected in suppressor and other similar effects. Obviously one could revert to direct comparison of the validity profiles of the raters but with 22 (or 23) points on each profile and the probability of complex patterns of emphasis interpretation by such inspection could become quite difficult.

It was primarily to enhance the identification of the content and differences among profiles that we turned to a factorial approach. It seemed obvious that the number of factors would be less than the number of predictors and that describing rater strategies by means of factor patterns would provide for easier interpretation than trait patterns. Hence we decided to do an inverse factor analysis using intercorrelations among the raters based upon each rater's profile of validity coefficients as a method to cluster the raters. Accordingly four separate factor analyses were carried out using the Hotelling principal axis method with Varimax rotation on the IBM 7094.

While using the results of the factor analyses to interpret the JAN outcomes, it occurred to us that we had perhaps developed an alternate method of clustering raters and capturing their strategies. Hence the remainder of this paper will consist of three major sections: (1) The outcome of the factor analyses; (2) a Profile of Factors (PROF) approach to the capturing of rater strategies; and (3) a comparison of the two strategy capturing methods.

Factor Analyses of Validity Coefficient Profiles

All four specialties yielded four identical factors. Three of the four specialties yielded a fifth factor. Factors were identified by selecting raters with relatively pure loadings on a single factor and computing their average validity coefficients for each trait. These average validity coefficients were then compared across factors to identify traits used most (having highest validities) by raters high on a given factor, the highest coefficient being underlined to indicate that the trait was to be used to identify that factor. In the factor descriptions to follow, the validities for each trait for a given factor have been averaged across all samples in which the trait occurred and in addition the ratio of cases in which it was underlined (was the highest value) is also reported. Thus a reported value of .550 (2/3) would mean that, for persons with a nearly pure loading on that factor, the average validity coefficient across the three

samples in which the trait was used amounted to .550 and that trait had the highest validity for these people on that factor in two of the three samples.

Factor A was identified as concern with *Leadership* on the basis of the following average validities:

Monitors work of subordinates	.722 (3/3)
Willing to delegate work	−.684 (2/2)
Willing to train men	−.682 (4/4)
Able to train others	−.677 (1/1)
Maintains discipline	.688 (4/4)
Willing to make decisions	.585 (2/2)
Supports his men	.574 (2/2)
Gets along with other	.572 (1/2)
Able to communicate	.550 (2/3)
Utilized opinions of subordinates	.478 (1/1)
Fair and impartial	.475 (2/2)
Informs his men	.464 (4/4)

This factor is seen to include items from the areas of both consideration and structuring by the leader with slightly higher validities for the structuring items.

Factor B was identified as concern with *Compliance* on the basis of the following average validities:

Punctual and dependable	.784 (4/4)
Care about details	.612 (2/2)
Care of property and equipment	.596 (3/4)
Willing to learn	.559 (2/4)
Neatness in work	.552 (2/2)
Safe work habits	.536 (3/4)
Good appearance	.453 (1/1)

This factor reflects high evaluation by the supervisor of the careful, neat, safe worker who is willing to do what he is told and is willing to learn how to do it. His person, his work, and the work he checks is always up to standard.

Factor C was identified as concern for *Job Skill and Knowledge* based on the following average validities:

Knows theory of aircraft systems	.866 (1/1)
Knows tech orders, etc.	.702 (4/4)
Plans and organizes work	.650 (3/4)
Able to train others	.649 (1/1)
Proficient in trouble shooting	.630 (2/2)
Able to communicate	.612 (3/3)
Speed and productivity	.602 (2/2)
Recognizes parts and equipment	.586 (3/3)
Utilizes diagrams, etc.	.563 (1/1)

Workers valued by supervisors high on this factor have ability, knowledge and skill and successfully apply them.

Factor D is identified as concern with *Motivation* on the basis of the following validities:

Plans and organizes work	.611 (1/4)
Willing to learn	.598 (2/4)
Job ingenuity	.548 (1/1)
Perseveres on the job	.523 (2/3)
Performs under pressure	.465 (2/4)
Physical Fitness	.266 (2/3)

Raters high on this factor show appreciation for the worker who prepares himself for the job, keeps himself in shape to perform well, and then sticks to the job even under adverse and trying conditions.

Factor E occurred in only three of the specialties, had relatively few high validities and these largely on items occurring in a single sample. It is, therefore, identified only tentatively as concern for *"Routine Thoroughness."* It is not, however, a very important factor as will become apparent later on. Its high validities were:

Careful in routine matters	.931 (1/1)
Gets along with others	.722 (1/1)
Thorough in inspections	.672 (1/1)
Satisfactory personal life	.498 (1/2)
Safe work habits	.427 (1/3)

That these factors do indeed explain most of the rater variance for all four specialties is shown in Table 1. The percentage of variance explained by each factor for each specialty was taken directly from the factor analysis output. From Table 1 we can conclude:

1. Supervision is judged to be a more important function for the skilled rather than the administrative type jobs and is apparently more important for the higher skill level.

2. Compliance is more evenly distributed in emphasis than are any of the other factors and in general it is not related to job level.

3. Job knowledge and skill are judged to be more important for the lower level skill job and for administrative job B.

4. The total importance of motivation is fairly equal for the two types of jobs but is much more unequally valued at the two versions of the administrative task while it is of more nearly equal importance for both levels of the skilled tasks.

5. Factor E makes relatively little contribution for any of the specialties.

6. The first four factors explain over 90 per cent of rater behavior in all four specialties.

TABLE 1

Proportion of Rater Variance Explained by Each Factor for Each of Four Different Specialties

Factor	Specialty: Lower Skill	Higher Skill	Admin. A	Admin. B
A. Supervision	42.34	47.34	25.88	29.81
B. Compliance	18.26	21.79	30.70	22.53
C. Job Knowledge and Skill	18.26	9.59	14.76	34.99
D. Motivation	12.39	13.77	18.82	2.44
E. Persistance (Routine)	1.60	1.74	—	3.30
Totals:	94.02	94.23	90.16	93.07

The Profile of Factors Method (PROF) for Capturing Rater Strategies

To implement the factor profile approach as a method for capturing strategies of raters it is necessary to redefine operationally what we mean by capturing a strategy. While JAN used the ability to predict what the rater did with the stimuli furnished him, PROF will define the capturing of a strategy as the identification of the factors (criteria) used by the raters and the relative importance assigned to each such factor. Raters with highly similar factor profiles will be judged to have the same strategies. While admittedly the similarity of profiles is hard to judge, some simplifying assumptions can at least make the task relatively objective.

In the present application of the method we decided to consider primarily three levels of rater involvement.

1. If a rater had a loading of .50 or more, i.e., had 25 per cent or more of his behavior controlled, on a given factor we could consider this a major part of his strategy. Such major components will be identified by capital letters.

2. If a rater had a loading between .30 and .49; i.e., 9 per cent up to 24.99 per cent of his behavior accounted for, on a given factor we could consider this a minor but still important part of his strategy. Such minor components will be identified by lower case letters.

3. Loading of less than .30 will, in general, be regarded as insufficient and be represented by a dot.

Thus raters with factor loading of 67, 19, 43, 07, —02 would be identified as having a strategy of (A.c..), while a rater with loadings of 03, 54, 63, 32, —09 would possess a strategy of (.B C d .). Raters with identical letter patterns will then be classified as having the same strategy. The nature of each strategy will be already predetermined by the letter pattern when the letters stand for the previously identified factors and the raters' level of usage.

Certain considerations led to these standards. For example, if a person is to give *major* consideration to several factors, we must define a major factor low enough to permit such a pattern to emerge. Thus with five factors we felt that 25 per cent might be an appropriate level. This will allow three (or maximally four) of the factors but *not all* of them to be considered as major foci of interest. With fewer factors the level might be set higher, while with more factors it would have to be reduced. Again at the lower level with five factors and average communalities around .90, if each factor was given its maximum usage each factor would account for 18 per cent of the variance. Thus the 9 per cent level (loading of 30) represents approximately one-half of the maximum possible strength under these conditions. Difference in number of factors and average communalities would probably necessitate some other standard. The important thing is to adopt some reasonable standard and then stick to it in order to keep the process as objective as possible.

Finally, in order to make the method conform to the JAN technique in so far as was possible, some attention was paid, in certain cases, to the distribution of the remaining loadings of less than .30. If a given rater has "insignificant" loadings which are all relatively high (+ .19, .23, .28 for example) he will tend to act more like raters with "significant" loadings on those factors than will a person who has all relatively low junk loadings (.04, — .02, .09) or one who has all relatively high (but insignificant) negative loadings (say —.21, —.24, —.18). Actually this type of consideration turned out to be important only in the skill specialties and only when a single significant high loading on Factor A was present.

TABLE 2

Strategies Captured for Four Job Specialties

PROF CODE	Lower Skill Specialty: Rater	Lower Skill Specialty: Strategy	Higher Skill Specialty: Rater	Higher Skill Specialty: Strategy	Admin. Specialty A: Rater	Admin. Specialty A: Strategy	Admin. Specialty B: Rater	Admin. Specialty B: Strategy
A – – – – –	35:	82 −14 −24 −23 −23						
A · · · ·	14:	93 14 17 07 07	14:	89 −10 20 27 −09	6:	91 18 −01 15		
	23:	98 01 05 05 −01	15:	94 18 −21 01 −10	24:	95 11 08 03		
			26:	97 06 −21 −05 −04				
			34:	98 13 −05 03 00				
			43:	96 08 04 −05 −02				
A + + + + +	1:	85 19 20 28 21	1:	90 26 04 28 14				
	7:	94 13 23 15 02	3:	91 09 26 25 07				
	17:	98 25 09 07 −15	11:	92 08 19 22 10				
	44:	91 13 23 23 −06	30:	92 23 04 21 10				
			36:	91 17 27 24 01				
			44:	90 28 18 20 04				
			46:	95 08 17 13 14				
			49:	99 03 04 14 02				
a · · · ·					5:	43 00 −01 09		
Ab · · ·	22:	90 64 −01 −10 −12	10:	67 51 22 22 −02	1:	84 36 12 13	10:	79 59 24 17 14
			12:	67 41 −08 13 −12				
			38:	90 52 −06 24 −05				
			39:	88 57 25 24 −02				
			45:	74 49 13 28 11				
AB · · ·	24:	62 74 −09 10 −11	5:	52 77 09 18 04	7:	59 67 25 14		
			42:	62 76 14 17 −05	10:	72 58 10 24		
			48:	78 66 02 23 05	17:	50 67 25 25		

TABLE 2—*Continued*

PROF CODE	Lower Skill Specialty						Higher Skill Specialty						Admin. Specialty A					Admin. Specialty B					
	Rater	Strategy					Rater	Strategy					Rater	Strategy				Rater	Strategy				
aB···	31:	87	85	27	13	06							32:	44	[illegible]	12	16	1:	86	87	15	12	07
																		34:	42	58	22	01	−16
																		41:	40	84	12	03	13
																		16:	86	23	89	10	07
																		36:	96	−18	82	08	03
A·c··	19:	85	05	84	28	18	21:	86	08	88	27	01											
	28:	83	07	45	24	00	27:	87	02	41	02	−13											
	40:	89	20	81	09	08																	
	41:	80	14	45	19	03																	
A·C··	3:	76	11	56	26	09							31:	[illegible]	14	[illegible]	−05	11:	68	29	66	−02	03
	18:	69	01	69	13	02												15:	51	05	84	−02	01
																		23:	77	27	58	−05	17
																		32:	65	03	67	00	07
																		38:	62	11	84	−03	22
																		30:	82	29	83	01	03
																		35:	41	01	77	19	16
a·C··	11:	49	20	74	20	06																	
	13:	91	06	83	23	07																	
A··C··													26:	[illegible]	28	−[illegible]	16						
A··d·	43:	86	29	22	[illegible]	07	50:	67	24	22	[illegible]	04											
A··D·							23:	[illegible]	24	16	[illegible]	04	25:	[illegible]	26	01	[illegible]						
a···E																							
Abc··	5:	87	44	49	02	05	25:	[illegible]	[illegible]	[illegible]	28	−02						44:	[illegible]	16	−20	04	[illegible]
	10:	48	86	44	29	−04	31:	[illegible]	[illegible]	[illegible]	−16	18						21:	[illegible]	[illegible]	[illegible]	08	12
	16:	84	88	47	15	02												49:	77	[illegible]	45	13	15
	29:	75	88	45	17	07																	
	30:	82	80	50	25	00																	
ABc··	4:	72	58	24	16	03	19:	[illegible]	67	[illegible]	28	13	4:	[illegible]	[illegible]	[illegible]	09	14:	72	58	79	07	18
													48:	[illegible]	[illegible]	[illegible]	26	25:	61	83	65	12	22
																		33:	56	62	48	11	23

AbC··				9: 62 41 58 27 12
				13: 59 37 66 15 02
				19: 67 44 52 11 12
				28: 61 39 65 06 07
				29: 60 34 68 13 16
				38: 60 38 65 19 11
				39: 67 39 60 15 −07
				42: 58 30 65 16 17
				45: 58 34 65 16 17
				48: 56 36 66 24 05
ABC··			13: 58 51 [illegible] 10	2: 50 57 61 23 01
				6: 50 48 51 03 12
				12: 57 57 60 01 02
aBC:	39: 43 61 62 20 06			37: 61 51 55 −01 13
				17: 45 60 52 02 28
				43: 42 65 53 05 07
abC··				46: 48 52 65 01 19
				18: 49 35 77 −04 16
Ab·d·	15: 69 59 20 45 −01	6: 71 49 08 39 −20	19: 71 45 29 37	24: 48 36 76 24 −16
Ab·D·			23: 76 42 −08 30	
			15: 59 37 05 64	
AB·d·		20: 53 59 03 48 −02	27: 59 71 15 57	
		28: 64 66 −09 63 02	28: 64 57 18 35	
		32: 51 61 29 41 −03	35: 62 54 17 38	
			38: 63 67 26 38	
aB·D·	37: 46 64 14 62 01		12: 43 59 29 63	
aB·d·			36: 33 77 27 46	
			41: 35 81 15 49	

TABLE 2—*Continued*

PROF CODE	Lower Skill Specialty						Higher Skill Specialty						Admin. Specialty A					Admin. Specialty B					
	Rater	Strategy					Rater	Strategy					Rater	Strategy				Rater	Strategy				
ab·D·	49:	47	52	11	73	11																	
aB··E							16:	41	62	−01	20	58											
A·cd·	2:	75	−01	46	39	−03																	
	12:	79	29	31	36	−05																	
A·Cd·	6:	72	27	51	30	04	4:	51	10	[illegible]	62	−12											
	20:	54	24	72	32	01																	
	27:	58	18	54	45	09																	
	36:	73	12	51	41	08																	
	42:	61	−12	56	37	10																	
A·cD·							22:	50	18	[illegible]	72	−03											
a·Cd·							29:	42	01	[illegible]	[illegible]	13											
A·c·e																							
a·C·e																		20:	73	25	47	08	30
Abcd·																		22:	40	26	66	−03	61
ABcd	8:	58	62	34	33	07							16:	62	37	30	49						
	25:	56	62	38	32	06							11:	58	52	44	40						
													20:	57	52	40	44						
													22:	57	60	38	32						
ABcD·													47:	52	60	42	47						
aBcd							40:	50	63	57	56	05											
							17:	42	65	34	41	−03	33:	45	62	45	40						
													34:	48	61	33	48						
													40:	37	74	39	30						
AbcD·	32:	60	45	35	56	68							45:	34	65	42	47						
													8:	59	34	37	55						
aBcD·													50:	57	47	39	56						
													18:	39	56	42	54						
													46:	49	52	51	56						

abCD·	21:	57	44	57	50	06							14:	53	36	57	60						
abcD·	46:	48	35	32	60	17	24:	36	47	43	63	−05						4:	45	47	42	56	08
·B···	26:	03	97	10	16	00	8:	−19	91	10	18	28	2:	−02	91	14	17	3:	−02	97	02	−06	10
							13:	−04	95	00	02	−15	32:	09	94	17	19						
							33:	17	96	−04	18	15	37:	24	91	−14	−13						
							35:	22	93	08	02	−14											
·BC··	47:	06	60	62	06	03												26:	16	64	65	11	25
																		31:	11	82	56	08	02
																		40:	24	52	78	08	01
·bC··	38:	07	90	91	11	00												5:	09	41	89	13	−01
																		8:	01	35	89	−02	−03
·B·d·	50:	15	91	18	50	00	18:	12	89	−06	39	−06											
·B·D·	34:	11	68	20	68	00	9:	11	62	07	65	−04											
·b·D·	33:	−01	88	23	86	−04							29:	22	84	28	83						
													49:	00	80	27	86						
·Bcd·													3:	24	77	82	47						
·BCd·							7:	02	54	58	49	14						27:	29	70	58	38	02
																		47:	28	55	65	87	00
·BCD·													44:	20	55	61	56						
·bCd·							41:	−22	89	78	82	−14											
·bcD·													43:	28	47	40	84						
·BcDe	48:	14	62	38	56	82																	
·B·dE	45:	−08	61	11	45	62																	
··C··	9:	17	18	93	08	00	2:	10	18	94	16	−01						7:	16	04	90	18	23
··Cd·													21:	05	27	90	80						
													42:	02	12	90	40						
··CD·													39:	−05	20	76	59						
···D·							37:	10	24	22	93	06	9:	25	−05	15	89						
							47:	19	18	23	93	03											

Table 2 shows the result of the application of the above standards to the four samples. In all, 58 separate strategies emerged. Not all of the strategies, however, occurred in any given sample. Actual count within the separate specialties indicated the following numbers of strategies:

Lower Skill	31
Higher Skill	25
Administrative A	29
Administrative B	20

We seem to have a tendency for more distinct strategies at the lower job level in the skill specialty and for the skilled as opposed to the administrative specialties.

Further discussion of the PROF results will be delayed until comparison is made with JAN outcome in the next section.

Comparison of PROF and JAN

The JAN technique had also been applied to all four samples of raters, the results for one specialty have already been reported in detail in our earlier article (Naylor and Wherry, 1965). It will be recalled that JAN had a built-in F test designed to indicate when combination of raters or rater groups into more comprehensive strategy patterns should stop. The PROF method simply combined all raters with the same upper case, lower case, and dot letter pattern. Our first comparison, therefore, will be made on the basis of the number of strategies reported by the two approaches. These were:

	Skill Specialties		Admin. Specialties	
	Lower	Higher	A	B
JAN	29	25	25	23
PROF	31	25	29	20

These results display considerable similarity. Both methods indicate more strategies for the Skilled than for Administrative Specialties and more strategies for lower level than for higher level within the skill specialty. Absolute differences in frequency are also quite small (2, 0, 4, 3) in all cases.

A second basis of comparison might be the variability in frequency of raters as they are distributed over the various strategies discovered. These comparisons are made in Table 3.

TABLE 3

Number of Raters in Strategies Captured by JAN and PROF

No. of Raters in Strategy	Proficiency: Skilled, Lower JAN	Skilled, Lower PROF	Skilled, Higher JAN	Skilled, Higher PROF	Administrative, A JAN	Administrative, A PROF	Administrative, B JAN	Administrative, B PROF
9							2	1
8			1	1				
7					1			
6	1							
5	1	2	2	2	3			1
4	1	2		1		3	1	1
3	1		3	3	2	2	1	5
2	7	5	4	2	3	8	6	5
1	18	22	15	16	16	16	13	7
Totals:	29	31	25	25	25	29	23	20

While the pattern of distribution is not identical for any of the specialties it appears that there is more similarity for the skilled than for the administrative specialties. Both methods agree that the strategies with the greatest consensus among raters occur for the higher Skill specialty and the administrative B specialty which show the only strategies with as many as eight or more raters in agreement.

Another basis of comparison would be in the factorial complexity of the strategies isolated. The number of strategies involving various degrees of complexity are shown in Table 4. Once again the two methods tend to show highly similar results with PROF showing a slight tendency to discover more complex patterns.

TABLE 4

Complexity of Strategies Captured by JAN and PROF

Number of Major and/or Minor Factors	Proficiency: Skilled, Lower JAN	Skilled, Lower PROF	Skilled, Higher JAN	Skilled, Higher PROF	Administrative, A JAN	Administrative, A PROF	Administrative, B JAN	Administrative, B PROF
4	4	5	1	3	5	6	1	1
3	8	9	9	10	6	10	9	9
2	12	12	6	7	7	9	11	8
1	5	5	9	5	7	4	2	2
Totals:	29	31	25	25	25	29	23	20

Since comparisons of JAN and PROF, up to this point, have dealt only with average factor loadings in the JAN strategies, it would still be possible for strategies to have the same "pattern" and even the same number of raters and still differ in the actual raters who were identified as belonging to the strategy. This final test—consisting of rater assignment is dealt with in Tables 5a, 5b, 5c, and 5d. These tables were constructed by using the PROF strategies and then extending the number of JAN clusters until the

TABLE 5a

Comparison of Strategies Captured by JAN and PROF for Lower Level Skill Specialty

PROF CODE	PROF	JAN
A ----	35	35
A ····	14, 28	1, 14, 23
A ++++	1, 7, 17, 44	7, 17, 30, 43, 44
Ab ···	22	22
AB ···	24	24
aB ···	31	31
A·c ··	19, 28, 40, 41	40
A·C ··	3, 18	3, 6, 18, 20, 28, 41
a·C ··	11, 13	11, 13
A··d·	43	
Abc··	5, 10, 16, 29, 30	4, 5, 8, 16, 25, 29, 10
ABc··	4	
aBC	39	39
Ab·d·	15	12, 15
aB·D·	37	37
ab·D·	49	49
A·cd	2, 12	2, 19, 27, 36
A·Cd·	6, 20, 27, 36, 42	42
ABcd	8, 25	
AbcD	32	32, 46
abCD	21	21
abcD	46	
·B···	26	26
·Bc··	47	47
·bC··	38	38
·B·d·	50	50
·B·D·	34	34
·b·D·	33	33
·BcDe	48	48
·B·dE	45	45
··C··	9	9
Number of Strategies:	31	33

TABLE 5b

Comparison of Strategies Captured by JAN and PROF for Higher Level Skill Specialty

PROF CODE	PROF	JAN
A····	14, 15, 26, 34, 43	14, 15, 26, 34, 38, 43
A++++	1, 3, 11, 30, 36, 44, 46, 49	3, 11, 36, 46, 49, 50, 1, 30, 44
Ab···	10, 12, 38, 39, 45	10, 19, 25, 12, 45, 48
AB···	5, 42, 48	5, 42
A·c··	21, 27	27, 39, 21
A··d·	50	
A··D·	23	22, 23
Abc··	25, 31	31
ABc··	19	
Ab·d·	6	6
AB·d·	20, 28, 32	28, 20, 32
aB··E	16	16
A·cd·	4	4
A·cD·	22	
a·Cd	29	29
ABcD·	40	40
aBcd	17	17
abcD	24	24
·B···	8, 13, 33, 35	8, 13, 33, 35
·B·d·	18	18
·B·D·	9	9
·BCd·	7	7
·bCd·	41	41
··C··	2	2
···D·	37, 47	37, 47
Number of Strategies:	25	32

greatest agreement between the two systems was achieved. As will be seen by inspecting these tables this resulted in using 33, 32, 33, and 33 JAN clusters in those tables.

To test the outcomes for agreement between the two systems, two types of overlap coefficients were computed. Identical raters in clusters with the same pattern were underlined and the proportion of common raters was divided by 50 (total number of raters) to get a Rater Overlap Coefficient. In addition the number of identically captured strategies, divided by the total number of different PROF strategies for the specialty is computed and called a Strategy Over-

lap Coefficient. The overlap coefficients based upon Tables 5a, 5b, 5c, and 5d were:

	Overlap Coefficients Rater	Overlap Coefficients Strategy
Lower Skill	.70	.87
Higher Skill	.84	.88
Lower Administrative	.70	.79
Higher Administrative	.84	.95

These coefficients indicate that the two methods arrived at essentially similar but not identical results.

As a matter of fact several of the non-agreements are more ap-

TABLE 5c

Comparison of Strategies Captured by PROF and JAN for Administrative Specialty A

PROF CODE	PROF	JAN
A····	6, 24	24, 6
a····	5	5
Ab···	1	
AB···	7, 10, 17	1, 10, 28, 17
aB···	32	32
A·C··	31	31
A·–C··	26	26
A··D·	25	25
ABc··	4, 48	4, 13
ABC·	13	
Ab·d	19, 23	23, 19
Ab·D	15	15
AB·d	27, 28, 35, 38	35
AB·D	12	
aB·d	36, 41	36, 41
Abcd	16	
ABcd	11, 20, 22, 47	11, 16, 33, 20, 22, 27, 38, 48
aBcd	33, 34, 40, 45	40, 34, 7, 18, 45, 47, 50
AbcD	8, 50	8, 12, 46
aBcD	18, 46	
abCD	14	14, 43
·B··	2, 30, 37	2, 30, 37
·b·D	29, 49	49, 29
·Bcd	3	3
·BCD	44	44
·bcD	43	
··Cd	21, 42	21, 42
··CD	39	39
···D	9	9
Number of Strategies:	29	33

TABLE 5d

Comparison of Strategies Captured by PROF and JAN for Administrative Specialty B

PROF CODE	PROF	JAN
Ab···	10	10
aB···	1, 34, 41	34, 1, 41
A·c··	16, 36	36, 16, 49
A·C··	11, 15, 23, 32, 38	38, 11, 23
a·C··	30, 35	35, 15, 30
a···E	44	44
Abc··	21, 49	
ABc··	14, 25, 33	33, 14, 21, 25
AbC··	9, 13, 19, 28, 29, 38, 39, 42, 45, 48	45, 18, 48, 19, 46, 13, 28, 29, 32, 39, 50, 9, 42
ABC··	2, 6, 12, 37	2, 12, 6, 37, 43
aBC··	17, 43, 46	17
abC··	18, 24	24
A·c·e	20	20
aC·e	22	22
abcD·	4	4
·B···	3	3
·BC··	26, 31, 40	31, 26, 40
·bC··	5, 8	5, 8
·BCd·	27, 47	27, 47
··C··	7	7
Number of Strategies:	20	33

parent than real and are concerned with the "insignificant" loadings discussed earlier in setting up the PROF system. For example, consider the code [A. C..] for the lower skill specialty

PROF: [A·C··]

3	76	11	56	26	09
18	69	01	69	13	02

JAN: [33], Combined at 33 Strategies [A·C··]

3	76	11	56	26	09	
6	72	27	61	30	04	A·Cd·
18	69	01	69	13	02	
20	54	24	78	32	01	A·Cd·
28	63	07	45	24	00	A·c··
41	60	14	45	19	03	A·c··

It might be said that JAN has erroneously added two variables with PROF pattern A. Cd. (6, 20) and two from pattern [A. c..] (28, 41). Actually, however, 3 and 18 do have loadings of 26 and 13 respectively on factor D while 6 and 20 have loadings of only .30 and .32. Had the D loadings on 6 and 20 been of the order .03 lower (an insignificant change) the PROF method would also have

combined them. Again variables 28 and 41 have factor loading of .45 on factor C, and if these loadings had been .05 higher (a probably insignificant change) we too would have combined these variables with 3 and 18. Since the boundaries for the PROF approach will always be relatively arbitrary, and since the JAN approach utilizes all bases for predictive trends we would always have many such *slight* differences in grouping.

There were some differences between the systems, however, as noted above under the discussion of overlap. In order to maximize similarity between the two systems it was necessary to extend the number of uncombined JAN groups not only beyond the criterion levels but actually beyond the number of strategies proposed by PROF. This was necessary because JAN refused to combine some groups suggested by PROF while combining some raters who were obviously quite different from the PROF viewpoint. For example, in Table 3 we see that while PROF pattern .B... combines raters 8, 13, 33, and 35; JAN (at 32 strategies) has them all as separates. Their loadings are (from Table 2):

8	−19	61	10	18	28
13	−04	98	00	02	−15
33	17	96	−04	18	15
35	22	[illegible]	08	02	−14

JAN did combine 33 and 35 at step 27, did not add 8 until step 14, and did not add 13 until step 10 after erroneously combining rater 13 with rater 18. In the meantime, JAN was making some rather peculiar combinations such as:

Step 38	Rater 27:	77	08	41	02	−13
	Rater 39:	[illegible]	37	25	24	02,
Step 30	Rater 6:	71	46	08	55	−20
	Rater 17:	43	[illegible]	[illegible]	41	−03,
Step 22	Rater 4:	61	10	68	32	−12
	Rater 14:	[illegible]	−10	20	27	−09,
Step 15	Rater 16:	40	62	−01	20	68
	Rater 31:	76	46	34	−16	18,

and the peculiar combination referred to above:

Step 13	Rater 13:	−04	[illegible]	10	02	−15
	Rater 18:	12	[illegible]	−06	[illegible]	−06

In the light of the above it seems that PROF may be a somewhat better basis for combinations of raters than is JAN, if the strategies are to be meaningful and interpretable.

It should be pointed out that PROF is closely related to more formal measures of profile similarity such as the Cronbach D^2 (Cronbach and Gleser, 1947) and the Cattell r_p (Cattell, 1949). These two measures are both joint measures of two characteristics which serve to differentiate profiles, namely, shape and general level. The PROF letter patterns also reflect these same differentiating characteristics. The capital letter, lower case letter, and dot patterns are rather obvious reflections of profile shape since they reflect relatively high, medium, and low points respectively. The level rules for defining such labels (.50 or above equals capital letter, .30 to .49 equals lower case letter; and .29 or less equals a dot) plus the definition of a strategy as a group of raters with identical letter patterns is designed to assure level distances between corresponding profile points will not exceed certain maximum values. Thus, the PROF system assures that r_p coefficients computed among raters in the same PROF strategy will be relatively high (D^2 would be correspondingly low). The converse—that raters from different strategies could necessarily have low r_p values—is not necessarily true, however. For example, raters with patterns (AbC..) and (aBc..) would be correctly identified as nearly identical if the actual loading were (.55, .45, .52, .08, —.03) and (49, 53, 46, 04, —09). However, even a person using PROF would have recognized that these patterns were more alike than those for another rater whose letter pattern was, say, (..cD.). Since no sampling error statistic is available for difference between r_p values, the use of such coefficients would scarcely serve as a more objective basis for classification (i.e., deciding whether a given rater should or should not be included in strategy *X* or set up as a separate strategy).

Actually to assume that there are any true boundaries which sharply separate one rater from another is probably untenable since raters can exhibit all shades and degrees of difference and similarity along the several dimensions isolated. To isolate regions of relative density within the factor hyperspace and to intelligently identify these regions is probably all that is required. The exact setting of boundaries for these regions so as to include or exclude a few fringe cases is probably of little relative importance. We chose, therefore, to not actually compute the more formal measures of profile similarity or to incorporate them in the PROF process.

REFERENCES

Bottenberg, R. A. and Christal, R. E. An Iterative Technique for Clustering Criteria which retains Optimum Predictive Efficiency. *WADD. Technical Note* 61–30, 1961.

Cattell, R. B. r_p and Other Coefficients of Pattern Similarity. *Psychometrika,* 1949, 14, 279–298.

Cronbach, L. J. and Gleser, G. C. Assessing Similarity Profiles. *American Psychologist,* 1947, 2, 425.

Naylor, J. C. and Wherry, R. J. The Use of Simulated Stimuli and the JAN Technique to Capture and Cluster the Policies of Raters. Educational and Psychological Measurement, 1965, 25, 969–986.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

STANINE VALUES FOR RANKS FOR DIFFERENT NUMBERS OF THINGS RANKED

C. J. BARTLETT and HAROLD A. EDGERTON
Performance Research, Inc.[1]

When confronted with the need for comparing relative class standing in classes of different sizes, the concept of the Stanine appears to be a useful answer.

The Stanine is a one digit number and occupies only one column in an IBM card. Its use assumes that the distribution of marks for the various classes have the same mean and standard deviation, and are normally distributed.

Table 1 was designed to convert the relative class standing of individual members of classes into Stanine values. To find the Stanine for a student standing 13th in a class of 37:

1. Go to the row "Number Ranked" = 37.
2. Follow across that row to the column which includes the student's rank 13.
3. The table entry shows that all of the ranks 9–15 inclusive belong in Stanine 6.

The table has been constructed to provide Stanines for all possible rankings in classes varying in size from 10 to 100.

TABLE 1

Ranks Corresponding to Stanine Values for Different Numbers of Things Ranked

NUMBER RANKED	STANINES 1	2	3	4	5	6	7	8	9	NUMBER RANKED
10	..	10	9	7–8	5–6	3–4	2	1		10
11	..	11	9–10	8	5–7	4	2–3	1		11
12	..	12	10–11	8–9	6–7	4–5	2–3	1		12
13	13	..	11–12	9–10	6–8	4–5	2–3	..	1	13
14	14	..	12–13	9–11	7–8	4–6	2–3	..	1	14
15	15	14	13	10–12	7–9	4–6	3	2	1	15
16	16	15	13–14	11–12	7–10	5–6	3–4	2	1	16
17	17	16	14–15	11–13	8–10	5–7	3–4	2	1	17
18	18	17	15–16	12–14	8–11	5–7	3–4	2	1	18
19	19	18	16–17	12–15	9–11	5–8	3–4	2	1	19
20	20	19	16–18	13–15	9–12	6–8	3–5	2	1	20
21	21	20	17–19	14–16	9–13	6–8	3–5	2	1	21
22	22	21	18–20	14–17	10–13	6–9	3–5	2	1	22
23	23	22	19–21	15–18	10–14	6–9	3–5	2	1	23
24	24	22–23	20–21	15–19	11–14	6–10	4–5	2–3	1	24
25	25	23–24	20–22	16–19	11–15	7–10	4–6	2–3	1	25
26	26	24–25	21–23	17–20	11–16	7–10	4–6	2–3	1	26
27	27	25–26	22–24	17–21	12–16	7–11	4–6	2–3	1	27
28	28	26–27	23–25	18–22	12–17	7–11	4–6	2–3	1	28
29	29	27–28	23–26	18–22	13–17	8–12	4–7	2–3	1	29
30	30	28–29	24–27	19–23	13–18	8–12	4–7	2–3	1	30
31	31	29–30	25–28	20–24	13–19	8–12	4–7	2–3	1	31
32	32	30–31	26–29	20–25	14–19	8–13	4–7	2–3	1	32
33	33	30–31	27–30	21–26	14–20	8–13	4–7	2–3	1	33
34	34	31–33	27–30	21–26	15–20	9–14	5–8	2–4	1	34
35	35	32–34	28–31	22–27	15–21	9–14	5–8	2–4	1	35
36	36	33–35	29–32	23–28	15–22	9–14	5–8	2–4	1	36
37	37	34–36	30–33	23–29	16–22	9–15	5–8	2–4	1	37
38	37–38	35–36	30–34	24–29	16–23	10–15	5–9	3–4	1–2	38
39	38–39	36–37	31–35	24–30	17–23	10–16	5–9	3–4	1–2	39
40	39–40	37–38	32–36	25–31	17–24	10–16	5–9	3–4	1–2	40
41	40–41	38–39	33–37	26–32	17–25	10–16	5–9	3–4	1–2	41
42	41–42	39–40	33–38	26–32	18–25	11–17	5–10	3–4	1–2	42
43	42–43	39–41	34–38	27–33	18–26	11–17	6–10	3–5	1–2	43
44	43–44	40–42	35–39	27–34	19–26	11–18	6–10	3–5	1–2	44
45	44–45	41–43	36–40	28–35	19–27	11–18	6–10	3–5	1–2	45
46	45–46	42–44	37–41	29–36	19–28	11–18	6–10	3–5	1–2	46
47	46–47	43–45	37–42	29–36	20–28	12–19	6–11	3–5	1–2	47
48	47–48	44–46	38–43	30–37	20–29	12–19	6–11	3–5	1–2	48
49	48–49	45–47	39–44	30–38	21–29	12–20	6–11	3–5	1–2	49
50	49–50	46–48	40–45	31–39	21–30	12–20	6–11	3–5	1–2	50
51	50–51	47–49	40–46	32–39	21–31	13–20	6–12	3–5	1–2	51
52	51–52	47–50	41–46	32–40	22–31	13–21	7–12	3–6	1–2	52
53	52–53	48–51	42–47	33–41	22–32	13–21	7–12	3–6	1–2	53
54	53–54	49–52	43–48	33–42	23–32	13–22	7–12	3–6	1–2	54
55	54–55	50–53	44–49	34–43	23–33	13–22	7–12	3–6	1–2	55
56	55–56	51–54	44–50	35–43	23–34	14–22	7–13	3–6	1–2	56
57	56–57	52–55	45–51	35–44	24–34	14–23	7–13	3–6	1–2	57
58	57–58	53–56	46–52	36–45	24–35	14–23	7–13	3–6	1–2	58

TABLE 1—*Continued*

NUMBER RANKED	STANINES 1	2	3	4	5	6	7	8	9	NUMBER RANKED
59	58–59	54–57	47–53	36–46	25–35	14–24	7–13	3–6	1–2	59
60	59–60	55–58	47–54	37–46	25–36	15–24	7–14	3–6	1–2	60
61	60–61	56–59	48–55	38–47	25–37	15–24	7–14	3–6	1–2	61
62	61–62	56–60	49–55	38–48	26–37	15–25	8–14	3–7	1–2	62
63	61–63	57–60	50–56	39–49	26–38	15–25	8–14	4–7	1–3	63
64	62–64	58–61	50–57	39–49	27–38	16–26	8–15	4–7	1–3	64
65	63–65	59–62	51–58	40–50	27–39	16–26	8–15	4–7	1–3	65
66	64–66	60–63	52–59	41–51	27–40	16–26	8–15	4–7	1–3	66
67	65–67	61–64	53–60	41–52	28–40	16–27	8–15	4–7	1–3	67
68	66–68	62–65	54–61	42–53	28–41	16–27	8–15	4–7	1–3	68
69	67–69	63–66	54–62	42–53	29–41	17–28	8–16	4–7	1–3	69
70	68–70	64–67	55–63	43–54	29–42	17–28	8–16	4–7	1–3	70
71	69–71	64–68	56–63	44–55	29–43	17–28	9–16	4–8	1–3	71
72	70–72	65–69	57–64	44–56	30–43	17–29	9–16	4–8	1–3	72
73	71–73	66–70	57–65	45–56	30–44	18–29	9–17	4–8	1–3	73
74	72–74	67–71	58–66	45–57	31–44	18–30	9–17	4–8	1–3	74
75	73–75	68–72	59–67	46–58	31–45	18–30	9–17	4–8	1–3	75
76	74–76	69–73	60–68	47–59	31–46	18–30	9–17	4–8	1–3	76
77	75–77	70–74	61–69	47–60	32–46	18–31	9–17	4–8	1–3	77
78	76–78	71–75	61–70	48–60	32–47	19–31	9–18	4–8	1–3	78
79	77–79	72–76	62–71	48–61	33–47	19–32	9–18	4–8	1–3	79
80	78–80	73–77	63–72	49–62	33–48	19–32	9–18	4–8	1–3	80
81	79–81	73–78	64–72	50–63	33–49	19–32	10–18	4–9	1–3	81
82	80–82	74–79	64–73	50–63	34–49	20–33	10–19	4–9	1–3	82
83	81–83	75–80	65–74	51–64	34–50	20–33	10–19	4–9	1–3	83
84	82–84	76–81	66–75	51–65	35–50	20–34	10–19	4–9	1–3	84
85	83–85	77–82	67–76	52–66	35–51	20–34	10–19	4–9	1–3	85
86	84–86	78–83	67–77	53–66	35–52	21–34	10–20	4–9	1–3	86
87	85–87	79–84	68–78	53–67	36–52	21–35	10–20	4–9	1–3	87
88	86–88	80–85	69–79	54–68	36–53	21–35	10–20	4–9	1–3	88
89	86–89	81–85	70–80	54–69	37–53	21–36	10–20	5–9	1–4	89
90	87–90	81–86	71–80	55–70	37–54	21–36	11–20	5–10	1–4	90
91	88–91	82–87	71–81	56–70	37–55	22–36	11–21	5–10	1–4	91
92	89–92	83–88	72–82	56–71	38–55	22–37	11–21	5–10	1–4	92
93	90–93	84–89	73–83	57–72	38–56	22–37	11–21	5–10	1–4	93
94	91–94	85–90	74–84	57–73	39–56	22–38	11–21	5–10	1–4	94
95	92–95	86–91	74–85	58–73	39–57	23–38	11–22	5–10	1–4	95
96	93–96	87–92	75–86	58–74	40–57	23–39	11–22	5–10	1–4	96
97	94–97	88–93	76–87	59–75	40–58	23–39	11–22	5–10	1–4	97
98	95–98	89–94	77–88	60–76	40–59	23–39	11–22	5–10	1–4	98
99	96–99	90–95	78–89	60–77	41–59	23–40	11–22	5–10	1–4	99
100	97–100	90–96	78–89	61–77	41–60	24–40	12–23	5–11	1–4	100

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

WEAK MEASUREMENTS VS. STRONG STATISTICS: AN EMPIRICAL CRITIQUE OF S. S. STEVENS' PROSCRIPTIONS ON STATISTICS[1, 2]

BELA O. BAKER
San Francisco State College
CURTIS D. HARDYCK
University of California Medical Center, San Francisco
and LEWIS F. PETRINOVICH
State University of New York at Stony Brook

The disagreement between those who belong to what Lubin (1962) called the "school of 'weak measurement' theorists" and those who belong to what might be called the school of "strong statistics" has persisted for a number of years with little apparent change of attitude on either side. Stevens, as the leading spokesman for the weak measurement school, has asserted (1951) and reasserted (1959, 1960) the view that measurement scales are models of object relationships and, for the most part, rather poor models which can lead one far astray from the truth if the scores they yield are added when they should only be counted. At least two current statistics texts intended for psychologists (Senders, 1958; Siegel, 1956) present this view as gospel.

Opposing this view, an assortment of statistically minded psychologists—e.g., Lord (1953), Burke (1953), Anderson (1961),

[1] This research was supported by research grants from the National Institutes of Health, U. S. Public Health Service (MH 07310) and the Research Committee, University of California Medical Center. Preliminary work was accomplished by a grant of free computer time by the Computer Center, University of California, Berkeley.

[2] We are grateful to Professor Jack Block, Professor Quinn McNemar, and Miss Mary Epling for their many helpful suggestions throughout this study. We are also indebted to Mrs. Eleanor Krasnow who developed and tested the computer programs used in this study.

McNemar (1962), and Hays (1963) have argued that statistics apply to numbers rather than to things and that the formal properties of measurement scales, as such, should have no influence on the choice of statistics. Savage (1957), a statistician, has supported this point of view, stating: "I know of no reason to limit statistical procedures to those involving arithmetic operations consistent with the scale properties of the observed quantities." In other words, a statistical test answers the question it is designed to answer whether measurement is weak or strong.

In his widely cited discussion of measurement, Stevens (1951) distinguished four classes of scales: Nominal, ordinal, interval, and ratio, and specified the arithmetic operations (and hence the statistics) which are permissible for each scale. Nominal scales consist simply of class names and can be treated only by counting operations and frequency statistics. Ordinal scales are developed by demonstrating that some objects have more of a particular quality than do other objects and representing numberically this order among objects. Lacking units, the numbers of an ordinal scale cannot be added, subtracted, multiplied, or divided, but they can be treated by order statistics such as the median or the rank-order correlation. Interval scales represent equal increments in the magnitudes of an object property by equal numerical increments. An increase of one unit in any region of an interval scale represents the same increment in the object property as does an increase of one unit in any other region of the scale. Scores from interval scales can be added and subtracted and hence such statistics as the mean, the standard deviation, and the product-moment correlation can be used. Ratio scales add a true zero point to equal intervals and can be multiplied, divided, and treated by subtle statistics which are of little concern to most psychologists.

Although Stevens develops his rationale for relating measurements and statistics almost exclusively in terms of descriptive statistics, he introduces the issue of hypothesis testing somewhat obliquely in his discussions of invariance of results under scale transformations (1951, 1959). He says, "The basic principle is this: Having measured a set of items by making numerical assignments in accordance with a set of rules, we are free to change the assignments by whatever group of transformations will preserve the empirical information in the scale. These transformations, depending

on which group they belong to, will upset some statistical measures and leave others unaffected (1959, p. 30)."

If parametric significance tests, such as *t* or F are used, the permissible transformations are linear. Only then will invariant results be found in comparing groups. An implication of this point of view, which is not made explicit by Stevens, is that if a scale is viewed as a model of object relationships, then any scale transformation is a transformation of those relationships. Hence the problem of invariance of results under scale transformations raises the following question: Can we make correct decisions about the nature of reality if we disregard the nature of the measurement scale when we apply statistical tests?

This aspect of Stevens' position has apparently been ignored by many of his critics. Anderson (1961) dismisses out of hand any restriction on the uses of *t* arising from the nature of the measurements to which it is applied but discusses the question of invariance of results under scale transformations seriously and at length before concluding that: "The practical problems of invariance or generality of results far transcend measurement scale typology" (p. 316).

The aspects of the problem as related to descriptive and inferential statistics are as follows: The problem for descriptive statistics as presented by Stevens (1951, 1959, 1960) concerns the relationship of the value of a particular statistic computed on obtained measurements to the value of the same statistic computed under conditions of perfect measurement. The argument is that the farther the measurement model departs from the underlying properties of the objects being measured, the less accurate the statistics. In other words, this aspect is concerned with precision of measurement.

In making statistical inferences, however, the issue is whether one will arrive at the same probability estimates from different types of measurement scales. Given the condition that a measurement scale may be a very poor model indeed of the properties of the objects under study, the question of the effect of the scale on the sampling distribution of a statistic remains unanswered. Where hypothesis testing is the issue, the appropriate question is: Do statistics computed on measures which are inaccurate descriptions of reality distribute differently than the same statistics computed under conditions of perfect measurement? If not, then a research

worker who has nothing better than an ordinal scale to work with may have to face the problem of more precise measurement for descriptive purposes, but at least the probability decisions he may make from his ordinal measurements will not be inappropriate for parametric statistical models.

In view of the importance of the issue raised by Stevens for users of statistics, it is surprising, as Lubin (1962) notes, that so few detailed discussions of the problem are available. If Stevens is correct, then psychologists should be disturbed about the state of their research literature. Since it can be safely asserted that most measurements in psychology yield scales which are somewhere between ordinal and interval scales, many psychologists may have been propagating fiction when they have made statistical inferences based on significance tests inappropriate to anything less than interval measurements. If Stevens' position is correct, it should be emphasized more intensively; if it is incorrect, something should be done to alleviate the lingering feelings of guilt that plague research workers who deliberately use statistics such as t on weak measurements.

A test of the issue would seem to require a comparison of the sampling distribution of a statistic computed under conditions of "perfect" measurement with the sampling distribution of the same statistic based on imperfect measurements. Since it is not possible to obtain such "perfect" measurements, this comparison is manifestly impossible. As noted above, however, Stevens has suggested that the main issue is that of invariance of results when measurement scales are transformed. Cast in these terms, the problem can be examined empirically. All that is required is that the sampling distribution of a statistic based on one set of scores be compared with the sampling distribution of the same statistic based on scores which are not "permissible" transformations of the first set. If Stevens is right, these sampling distributions should differ in some important way. If they do not, then the nature of the measurement scale is, within potentially determinable limits, an irrelevant consideration when one chooses an hypothesis testing statistic.

Method

The statistic selected for study was Student's t. Not only is this one of the most commonly used statistics in psychology but it also

has the advantage of having been demonstrated empirically to be relatively robust in the face of violations of the assumptions of normality and equality of variances (Norton, 1953; Boneau, 1960).

The first set of scores used—which will be referred to as criterion (unit-interval) scores—comprised the cardinal numbers from 1 to 30. Three populations of 1000 scores each were constructed by assigned frequencies to the unit-interval scores to approximate as closely as possible the expected frequencies for (1) a normal distribution, (2) a rectangular distribution, and (3) an exponential distribution ($f = 1 + 275^{(e - .25X)}$).

According to Stevens (1951), when one uses the mean or standard deviation the permissible score transformations are linear. And so, to evaluate this proscription, 35 non-linear transformations of the unit-interval scores were constructed. These fell into three subsets, each designed to simulate a common measurement problem in psychology. The rationale for this approach is that an investigator can almost always develop a measuring device that looks as if it yields scores with equal intervals. However, relations among the objects represented by these numbers may not be equal—this is, of course, Stevens' main concern—and consequently by producing non-linear transformations of cardinal numbers, a class of situations is produced directly analogous to the situation that obtains when a measurement scale correctly represents the order among objects but incorrectly represents the magnitude of differences between objects. A statistic such as t, it is argued, cannot be used with such a measuring device, since the operations of addition, subtraction, multiplication, and division are inappropriate—a condition which rather limits the investigator seeking statistical support for his conclusions.

The first set of transformations (1 through 15) was constructed to simulate the situation in which an ordinal scale is achieved but interval sizes vary randomly. This may often be the case when the sum of responses in a keyed direction on a personality inventory is used as a trait measure. The first score of each transformation was the number 1, to which was added a number selected at random within pre-set limits to produce the second score. Another random number was then added to the second score to produce the third number, continuing until 30 scores had been developed. With this technique, the intervals between successive scores could vary ran-

domly from 1 to a pre-selected maximum. In order to examine the effects of the magnitude of interval variations, three different maximum interval sizes were used in the first 15 transformations. Transformations 1 through 5 were constructed with a maximum of 2, allowing the largest interval to be twice as large as the smallest. Transformations 6 through 10 were constructed with a maximum of 10, and transformations 11 through 15 were constructed with a maximum of 25. Therefore, in transformations 11 through 15 one interval potentially could be 25 times as large as another. Figure 1 illustrates the relationship between the scores of transformation 7 and the unit-interval criterion scores. Each preselected maximum value was used to develop five transformations—a precaution which served to cancel out artifacts of sampling for a particular transformation.

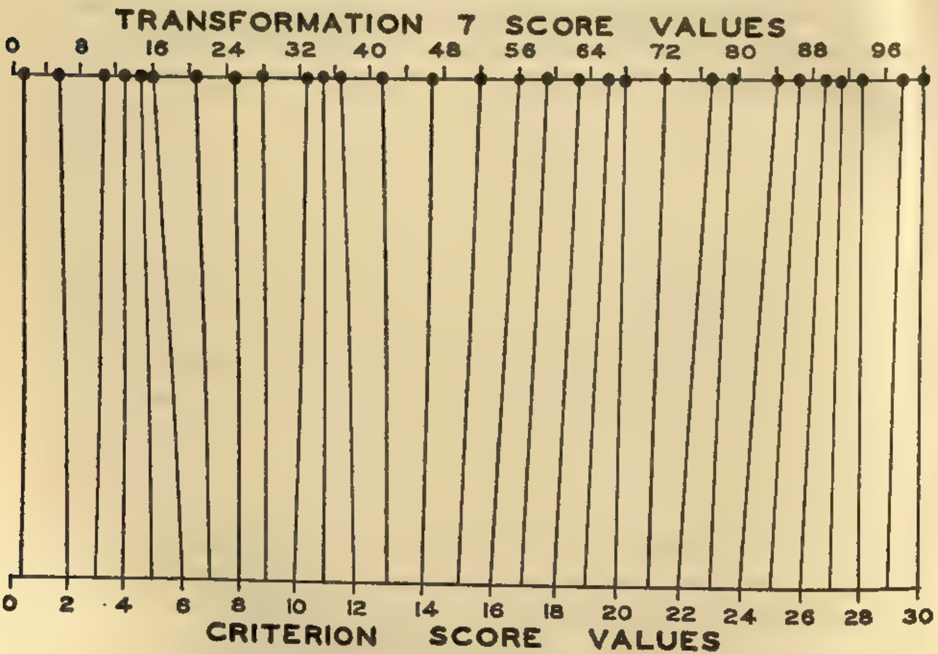


Figure 1. Relationship between the intervals of Transformation 7 and the unit intervals of the criterion scores.

The second set of transformations (16 through 25) was devised to simulate the situation which may often occur in the measurement of achievement and ability in which the magnitudes of trait

differences represented by intervals at the extremes of a scale may be greater than those represented by equal appearing intervals in the middle of the scale. Each transformation was produced by choosing a number from 1 to a relatively large maximum for the first interval, a number from 1 to a somewhat smaller maximum for the second interval, and so forth, progressively reducing the maximum until the center of the score array was reached, then increasing the maximum at the same rate to the end of the score array. For transformations 16–20, the series of maximum values decreased by one unit steps from a possible maximum of 15 at the beginning of the array to 1 at the center of the array, increasing again by one unit steps to a possible maximum of 15 again at the other end of the array. For transformations 21–25, the maximum was decreased by 3 unit steps from 45 to 3 at the center of the array and back to 45. The type of transformation produced by this procedure is illustrated in Figure 2.

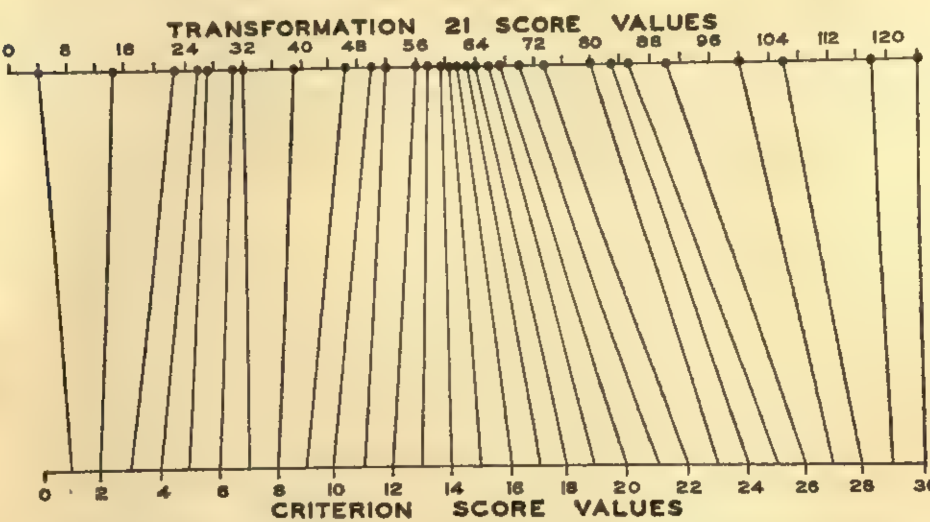


Figure 2. Relationship between the intervals of Transformation 21 and the unit intervals of the criterion scores.

For the third set of transformations (26 through 35), unit intervals were retained for scores ranging from 1 to 15, the remaining 15 intervals being varied randomly. This procedure crudely imitates a problem which may be present sometimes in social distance scales or in the Thurstone type scaling of attitude items. For trans-

formations 26–30, the maximum interval size above the center of the score array was 10. For transformations 31–35, the maximum was 25. Figure 3 illustrates this type of transformation.

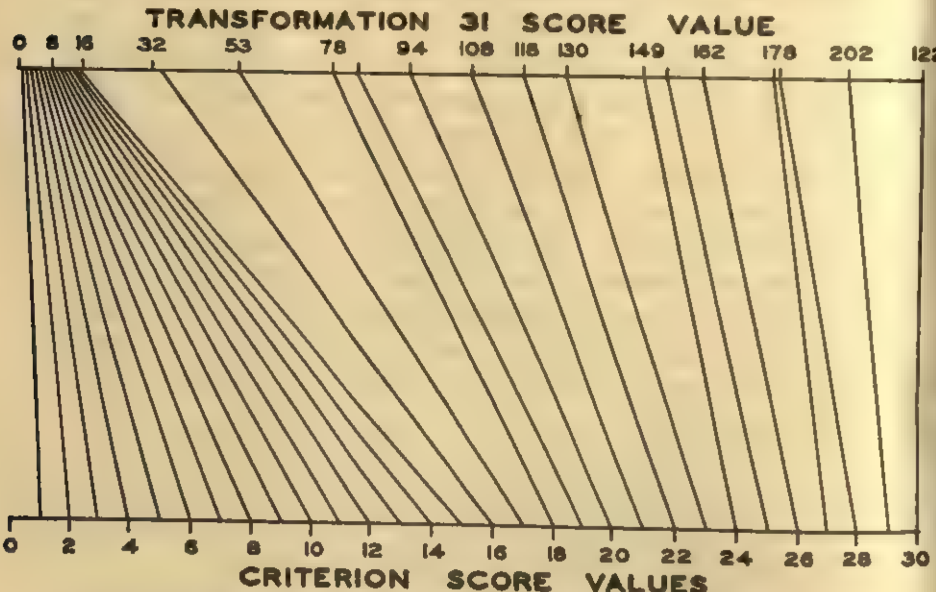


Figure 3. Relationship between the intervals of Transformation 31 and the unit intervals of the criterion scores.

For each of the populations, the unit-interval scores and the transformations were stored in the memory of an IBM 7090 digital computer as 36 scores for each of 1000 cases, which were identified by the numbers 1 through 1000. Random samples of cases were selected by means of a multiplying random number generator, with some additional features to insure complete randomness and to guard against cycling. The generator sequence was initiated by obtaining a 35 binary digit (bit) number from the clock of the computer. This number was forced to be odd by the insertion of a low-order bit. A tape containing the one million random digits produced by the RAND Corporation (RAND, 1955) was then referenced and a word of ten decimal digits read from this tape. The first half of the 10 digit word, R, was then transformed by the value 8R + 3. This value served as the initial multiplier for the number obtained from the computer clock. This multiplication produced a 70 bit number, the right 35 bits of which were ex-

tracted from the random number. The left bits of the latter were used to obtain an integer in the range 1 — 1000. This value determined the "case" to be selected from the population. The 35 bit random number was retained and multiplied by 8R + 3 to obtain the next value from the population.

The second half of the ten-digit number first read from the RAND tape was divided by the quantity (NA + NB)/2, where NA and NB are the respective sizes of the pairs of samples drawn. The remainder occurring from this division determined the number of times the value 8R + 3 was used as a multiplier in generating a random sequence. When the number of values generated equaled the value of the remainder, another ten digit record was read from the RAND tape and the sequence of generation repeated. When samples of size NA and NB had been drawn, t was computed using a standard computing formula.

A total of 36 t values were computed for each pair of samples drawn: One value for the unit-interval or criterion scores and one for each of the 35 transformations of the criterion. This is analogous to sampling from a pool of 1000 subjects with scores on 36 variables, the first variable representing a set of measurements with equal intervals and the remaining 35 variables representing measurement scales standing in varying non-linear relationships to the first set.

The following notation will be used throughout the paper:

N,R,E ....the type of distribution used: Normal, Rectangular, or Exponential
5,5; 15,15; 5,15 .....size of first and second samples
C ....the criterion set of values
T_n ....the nth transformation
T_{j-k} ....transformations j through k.

The sequence of the notation is as follows: N,15,15,T_{6-10}; a normal distribution with sample sizes of 15 and 15 for transformations 6 through 10.

The computations were summarized in three forms: (1) Contingency tables showing the relationship of the criterion t value to each transformation t value for the .01 and .05 significance levels were tabulated. These tables allow the determination of the difference in the percentage of t's reaching given significance levels for a particular transformation and for the criterion scores. (2) Fre-

quency distributions of all sets of t values were tabulated. Figure 4 shows the frequency distributions of t for N, 5, 5, C; N, 5, 5, T_5; N, 5, 5, T_{20}; and N, 5, 5, T_{35}.

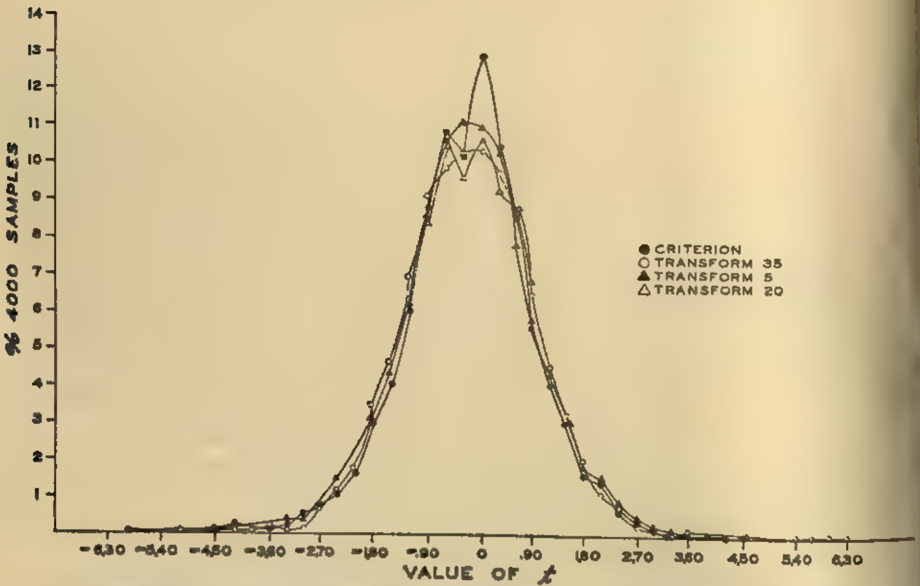


Figure 4. Empirical sampling distributions of t for 4000 pairs of samples with NA = NB = 5, from a normal distribution for the Criterion and Transformations 5, 20, and 35.

(3) Pearson correlation coefficients between the criterion t values and each of the transformation t values were calculated over the total number of sample pairs drawn for a given condition. The standard error of estimate was also computed. These statistics provide estimates of the degree to which t's based on transformed scores varied from t's based on unit-interval scores for the same pairs of samples.

Parts (1) and (2) of the computation summaries are directly relevant to the question of the effect of the measurement model on the sampling distribution of t. Part (3) is an attempt to represent the deviation of the various types of "inappropriate" measurement models from a true value (represented by the criterion).

As has been mentioned earlier, three types of distributions—normal, rectangular, and exponential, were used. Three variations in sample size were studied—NA = NB = 5; NA = NB = 15;

and NA = 5, NB = 15. These combinations are identical with those used by Boneau (1960) and permit the present results to be compared directly with his.

Results

As a first step, the distributions of t on the criterion measurements for all conditions were compared with the theoretical distributions for the appropriate degrees of freedom. Table 1 contains the percentage of t values falling in the 5 per cent and 1 per cent regions of the distribution for the criterion scores.

TABLE 1

Per cent of t's Based on Criterion Unit Interval Scores Falling in the 5% and 1% Regions of Rejection for 4000 Random Sampling Runs

	Sample Sizes					
	NA = NB = 5		NA = NB = 15		NA = 5, NB = 15	
Population Distribution	5% Level	1% Level	5% Level	1% Level	5% Level	1% Level
Normal	4.8	.8	5.4	1.0	5.3	1.1
Rectangular	5.4	1.6	4.6	.9	5.1	.8
Exponential	3.9	.9	5.1	1.0	4.2	.9

The deviations of the empirical distributions from the expected theoretical values for the normal curve are quite small. The results are very similar to those reported by Boneau (1960), including the underestimates for the exponential distribution. Somewhat surprisingly, the deviations are reduced only slightly from those reported by Boneau, despite the fact that the present results are based on 4000 t's as compared to Boneau's 1000.

The results for the first set of transformations, which were constructed to simulate a situation where intervals vary randomly throughout the range of the measuring instrument, are given in Table 2. Since the tabulation of results for individual transformations within and across sets T_{1-5}, T_{6-10}, and T_{11-15} revealed little variation, only mean values for all transformations are given in Table 2.[3] The mean value tabled for each transformation is based on 60,000 t's.

Examination of Table 2 indicates that random variations tend to have little effect on the number of t's falling in the 5 per cent and

[3] Complete tables are available on request from the authors.

TABLE 2

Per cent of t's Falling in the 5% and 1% Regions of Rejection When Interval Sizes Vary Randomly (4000 Samples Per Condition)

		5% Level		1% Level	
		Total %	% in Larger Tail	Total %	% in Larger Tail
N, 5, 5:	C	4.8	2.4	.8	.4
	T_{1-15}	4.8	2.4	.8	.4
N, 15, 15:	C	5.4	3.0	1.0	.5
	T_{1-15}	5.3	2.9	.9	.5
N, 5, 15:	C	5.3	2.7	1.1	.6
	T_{1-15}	5.0	2.6	1.1	.6
R, 5, 5:	C	5.4	3.0	1.6	.9
	T_{1-15}	5.3	3.0	1.6	.9
R, 15, 15:	C	4.6	2.6	.9	.5
	T_{1-15}	4.6	2.5	.9	.5
R, 5, 15:	C	5.1	2.7	.8	.7
	T_{1-15}	5.0	2.6	.9	.7
E, 5, 5:	C	3.9	2.0	.9	.5
	T_{1-15}	3.9	2.0	.8	.4
E, 15, 15:	C	5.1	2.6	1.0	.6
	T_{1-15}	5.2	2.6	1.0	.6
E, 5, 15:	C	4.2	3.8	.9	.9
	T_{1-15}	4.2	3.8	1.0	1.0

1 per cent regions of rejection. Columns (1) and (3) contain the total percentages for the 5 per cent and 1 per cent levels of the *t* distribution for all distributions and sample sizes. For the first group of transformations, the total percentages in the critical regions are very close to the theoretical expectation for a normal distribution. The largest discrepancy present is for E, 5, 5,—a discrepancy of only 1.1 per cent. Columns (2) and (4), which contain the percentage in the larger tail, show similar minimal variations. If one allows for the effects of sampling and takes the deviations of each transformation from the obtained percentages of the C distribution, the discrepancies become almost nonexistent. The largest deviation is .3 per cent for N, 5, 15. It is evident that random variations in interval sizes, regardless of the magnitude of those varia-

tions, have virtually no effect on the percentage of t's reaching conventional significance levels.

One condition reported in Table 2 (E, 5, 15) did result in an asymmetrical t distribution. However, as can be seen by examining the tabled values for E, 5, 15, the C distribution is equally asymmetrical. For E, 5, 15, the majority of the t values reaching the 5 per cent level and all of the t's at the 1 per cent level are in one tail of the distribution. The direction of the skewing is negative, indicating that where large differences between sample means occurred, the higher mean tended to be the mean of the smaller sample. On the basis of this finding, an experimenter would be ill-advised to use a one-tailed test when he is using samples of unequal sizes (at least if the sizes are of the magnitudes used in this study). However, it makes little difference whether there is an interval scale of measurement or not.

The results for the more irregular and extreme transformations (16 through 35) are presented in Table 3. Again, since there was little variation within sets only mean values for transformations 16–25 and transformations 26–35 are presented.

An inspection of Table 3 also permits the conclusion that the magnitude of variations in interval sizes has little effect on the t distribution. At the same time, it is apparent that t is affected more by these types of transformations than was the situation for simple random variation. However, the discrepancies are still far from extreme. In columns (1) and (3) the largest obtained discrepancy is 2.3 per cent for E, 15,15,T_{26-35} at the 5 per cent level. In columns (2) and (4), the largest discrepancy is again at the 5 per cent level for E, 15,15,T_{26-35}, a value of 1.1 per cent.

When compared to the 5.1 per cent of t's falling in the 5 per cent region for the E, 15, 15, C distribution, this discrepancy of 2.3 per cent seems rather large. However, it seems slight compared to the discrepancies obtained when more serious violations of the assumptions for the use of t are made. For example, Boneau (1960) reported 16 per cent of obtained t's at the 5 per cent level for samples of 5 and 15 drawn from normally distributed populations with unequal variances.

When Table 3 is examined for asymmetry, it is evident that the transformations in which the intervals in one half of a scale stand for substantially smaller variations in the objects being measured

TABLE 3

Per cent of t's Falling in the 5% and 1% Regions of Rejection When Interval Sizes Vary More in Some Regions of the Scale Than in Others (4000 Samples Per Condition)

		5% Level		1% Level	
		Total %	% in Larger Tail	Total %	% in Larger Tail
N, 5, 5:	C	4.8	2.4	.8	.4
	T_{16-25}	3.4	1.8	.4	.3
	T_{26-35}	4.2	2.2	.8	.4
N, 15, 15:	C	5.4	3.0	1.0	.5
	T_{16-25}	4.6	2.6	.6	.4
	T_{26-35}	5.1	2.8	.9	.5
N, 5, 15:	C	5.3	2.7	1.1	.6
	T_{16-25}	5.0	2.7	1.0	.6
	T_{26-35}	4.7	3.6	1.0	.9
R, 5, 5:	C	5.4	3.0	1.6	.9
	T_{16-25}	4.9	2.7	1.0	.6
	T_{26-35}	4.4	2.6	.8	.5
R, 15, 15:	C	4.6	2.6	.9	.5
	T_{16-25}	4.6	2.4	.8	.4
	T_{26-35}	4.9	2.8	.8	.4
R, 5, 15:	C	5.1	2.7	.8	.7
	T_{16-25}	4.9	2.6	.8	.6
	T_{26-35}	4.3	3.1	1.1	1.1
E, 5, 5:	C	3.9	2.0	.9	.5
	T_{16-25}	5.2	2.8	1.4	.7
	T_{26-35}	3.3	1.6	.7	.4
E, 15, 15:	C	5.1	2.6	1.0	.6
	T_{16-25}	5.4	2.7	1.2	.6
	T_{26-35}	2.8	1.4	.4	.3
E, 5, 15:	C	4.2	3.8	.9	.9
	T_{16-25}	4.6	3.4	.8	.7
	T_{26-35}	3.8	3.6	.6	.6

than do intervals in the other half of the scale—T_{26-35}—yield seriously skewed distributions of t for all conditions where unequal sample sizes are used. For E distributions, skewing is present for most of the transformations. These transformations provide the only situation where the nature of the scale transformation affected the sampling distribution of t to a more serious degree than could be attributable to the use of unequal sample sizes drawn from an

exponential distribution. Even for this condition the effect is quite small. For any real-life situation in which the possibility of such a measurement scale exists, an experimenter should be chary of using t to make a one-tailed test between means based on unequal sample N's. Fortunately, this problem occurs only rarely and when it does occur the use of equal sample sizes will minimize the distortion.

In reviewing the results presented so far, the following generalizations seem warranted:

1. The percentage of t's reaching the theoretical 5 per cent and 1 per cent levels of significance is not seriously affected by the use of non-equal interval measurements.[4]

2. To the extent that there is any influence of the scale transformation on the percentage of t's reaching theoretical significance levels, the influence is more marked when intervals in one broad region of a scale are larger than intervals in another region of the scale than it is when interval sizes vary randomly.

3. If an investigator has a measuring instrument which produces either an interval scale or an ordinal scale with randomly varied interval sizes, he can safely use t for statistical decisions under all circumstances examined in this study. The single exception is that t should not be used to do a one tailed test when samples of unequal size have been drawn from a badly skewed population.

4. If a measurement scale deviates from reality in such a fashion that the magnitude of trait differences represented by intervals at the extremes of the scale may be greater than those represented by equal-appearing intervals in the middle of the scale ($T_{16\text{–}25}$), it seems reasonably safe to use t. Unequal sample sizes can even be used if the population is symmetrical, but the proscriptions against using one-tailed tests for unequal sample sizes from exponential populations still apply.

5. If the scale is of the kind represented by the relationship between C and $T_{26\text{–}35}$ (in which inequality of units is present in one-half of the distribution only), it is still safe to use t, with a somewhat stricter limitation on the use of one-tailed tests. This arises

[4] It is possible that the effects of the scale transformations used in this study are actually due to changes in the shape of the distributions which the different transformations produced. However, if this is the case, the arguments presented regarding the insignificant effects of the nature of measurement scales on probability statements are strengthened even more.

from the finding that for all population distributions these transformations yielded skewed distributions of t when unequal sample sizes were used.

6. As a maximally conservative empirical set of rules for using t, the following restrictions would seem to be sufficient to compensate for almost any violation of assumptions investigated up to this time:

a. Have equal sample sizes.

b. Use a two-tailed test.

7. Returning to the question as originally formulated: Do statistics computed on a measurement scale which is at best a poor fit to reality distribute differently than the same statistics computed under conditions of perfect measurement? The answer is a firm "no," provided that the conditions of equal sample sizes and two-tailed tests are met. The research worker who has nothing better than an ordinal scale to work with may have an extremely poor fit to reality, but at least he will not be led into making incorrect probability estimates if he observes a few simple precautions.

As a final step, a different sort of analysis will be cited. The previous results and discussion related to one aspect of the measurement problem as posed by Stevens (1951); a second aspect remains. This concerns the accuracy of the descriptive statistics when the measurement model is a poor fit. Stevens has presented his point of view almost exclusively in terms of descriptive statistics and has tended to use illustrations from descriptive statistics to support his arguments. In the last analysis, this would seem to raise an epistemological question, since it is concerned with the relationship of measurement to a true value which cannot be known. However, evidence as to the correctness or incorrectness of the point of view can be examined from the data of the present study, even though the results are of no help in solving the problems faced by an experimenter who is wondering how to evaluate the validity and the precision of his measuring instrument.

The question of the accuracy of representation can be evaluated by defining the unit interval criterion t values as true measures and the values calculated on the various transformations as those obtained on a measurement model which misrepresents reality. Then the degree of relationship between the values of t calculated on specific samples for C and the values calculated on T_{1-35} can be ob-

tained. This is a correlational question and the results are reported in Table 4.

Columns (1), (3), and (5) contain for each of the distributions the correlations between values of t for each set of transformations and the corresponding values of t for the criterion. The correlations are impressively high. However, because of the broad range of values in the t distribution, the standard errors of estimate in columns (2), (4), and (6) are more informative statistics.

TABLE 4

Mean[a] Correlation Coefficients and Standard Errors of Estimate for the Prediction of t's Based on Transformed Scores from t's Based on Criterion Unit-interval Scores

Population Distribution		NA = NB = 5		NA = NB = 15		NA = 5, NB = 15	
		Mean r (1)	Mean $s_{y\cdot x}$ (2)	Mean r (3)	Mean $s_{y\cdot x}$ (4)	Mean r (5)	Mean $s_{y\cdot x}$ (6)
N:	T_{1-5}	.997	.089	.997	.082	.997	.084
	T_{6-10}	.996	.111	.995	.100	.995	.104
	T_{11-15}	.992	.146	.991	.138	.991	.142
	T_{16-20}	.975	.244	.966	.265	.970	.260
	T_{21-25}	.968	.271	.964	.274	.966	.278
	T_{26-30}	.935	.401	.933	.380	.933	.386
	T_{31-35}	.914	.462	.911	.434	.912	.439
R:	T_{1-5}	.999	.056	.988	.048	.999	.033
	T_{6-10}	.996	.094	.996	.081	.996	.084
	T_{11-15}	.994	.117	.994	.088	.988	.104
	T_{16-20}	.973	.256	.973	.231	.973	.233
	T_{21-25}	.973	.258	.975	.227	.976	.225
	T_{26-30}	.948	.368	.943	.339	.944	.348
	T_{31-35}	.927	.430	.922	.394	.924	.404
E:	T_{1-5}	.994	.121	.993	.113	.992	.117
	T_{6-10}	.992	.138	.992	.126	.992	.129
	T_{11-15}	.984	.199	.985	.181	.983	.189
	T_{16-20}	.970	.283	.946	.342	.951	.324
	T_{21-25}	.963	.313	.953	.314	.954	.309
	T_{26-30}	.981	.218	.930	.382	.940	.325
	T_{31-35}	.966	.288	.885	.483	.922	.405

[a] Median values do not differ until the third decimal place for the majority of transformations.

Several points can be noted in connection with Table 4: There is a regular progression in the size of the standard errors of estimate across the sets of transformations used, such that they are smallest for T_{1-15}, and largest for T_{26-35}. These standard errors also become

larger as the magnitude of variations in interval size increases, but this is less striking than the differences among types of transformations. Variations in sample sizes and in the shape of the population distribution do not seem to have much influence on the standard errors of estimate; consequently these results seem to show a specific influence of scale transformations on the values of t. The correspondence between values of t based on the criterion unit interval scores and values of t based on transformations decreases regularly and dramatically—from standard errors of estimate on the order of .08 to standard errors of estimate on the order of .45—as the departure from linear transformations becomes more extreme. Here, then, is a finding consistent with Stevens' expectations: The value of t determined for a comparison of samples of non-interval scores does tend to be different from the value of t based on interval scores for the same samples and the discrepancy tends to become greater as the departure from equal intervals is more marked.

In conclusion, the views presented by Stevens (1951, 1959, 1960) and by advocates of his position such as Senders (1958), Siegel (1956), and Stake (1960) state that, when one uses t, the measurement model should have equal intervals representing linear transformations of the magnitudes of the characteristics being measured, or the statistic will be "upset." This view may be correct if one considers single specific determinations of a statistic in a descriptive sense—this seems to be the significance of the standard errors of estimate reported in Table 4—but it is incorrect when applied to the problem of statistical inference.

The present findings indicate that strong statistics such as the t test are more than adequate to cope with weak measurements—and, with some minor reservations, probabilities estimated from the t distribution are little affected by the kind of measurement scale used.

REFERENCES

Anderson, N. H. Scales and Statistics: Parametric and Nonparametric. *Psychological Bulletin*, 1961, 58, 305–316.

Boneau, C. A. The Effects of Violations of Assumptions Underlying the t Test. *Psychological Bulletin*, 1960, 57, 49–64.

Burke, C. J. Additive Scales and Statistics. *Psychological Review*, 1953, 60, 73–75.

Hays, W. L. *Statistics for Psychologists.* New York: Holt, Rinehart and Winston, 1963.

Lord, F. M. On the Statistical Treatment of Football Numbers. *American Psychologist*, 1953, 8, 750–751.

Lubin, A. Statistics. In *Annual Review of Psychology.* Palo Alto, Calif.: Stanford University Press, 1962.

McNemar, Q. *Psychological Statistics,* 3rd ed. New York: Wiley, 1962.

Norton, D. W. An Empirical Investigation of Some Effects of Non-Normality and Heterogeneity on the F-Distribution. Unpublished Doctoral Dissertation, State University of Iowa, 1952. Cited in E. F. Lindquist, *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton-Mifflin, 1953.

RAND Corporation. *A Million Random Digits.* New York: The Free Press of Glencoe, 1955.

Savage, I. R. Non-parametric Statistics. *Journal of the American Statistical Association,* 1957, 52, 331–344.

Senders, V. L. *Measurement and Statistics.* London: Oxford University Press, 1958.

Siegel, S. *Nonparametric Statistics.* New York: McGraw-Hill, 1956.

Stake, R. E. Review of *Elementary Statistics* by P. G. Hoel. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 871–873.

Stevens, S. S. Mathematics, Measurement and Psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* New York: Wiley, 1951.

Stevens, S. S. Measurement, Psychophysics and Utility. In Churchman, G. W., and Ratoosh, P. (Eds.), *Measurement: Definitions and Theories.* New York: Wiley, 1959.

Stevens, S. S. Review of *Statistical Theory* by Lancelot Hogben. *Contemporary Psychology,* 1960, 5, 273–276.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

POWER OF THE ANALYSIS OF VARIANCE OF INDEPENDENT GROUPS ON NON-NORMAL AND NORMALLY TRANSFORMED DATA

PAUL A. GAMES and PATRICK A. LUCAS[1]
Ohio University

The science of data analysis, as described by Tukey (1962), (in contrast to statistics), has a goal closely tied to the real world: to derive certain conclusions about the "state of the world" from a sample of evidence. Statistics, on the other hand, is a mathematical discipline and can't be evaluated by the same empirical considerations. Consequently, although data analysis finds statistics very useful, it is helpful to maintain a distinction when approaching the practical problems considered here. Two aspects of data analysis are relevant to this paper: (1) the choice of the best statistical test for the analysis desired and (2) an evaluation of how the test will be affected by deviations of the data from the assumptions of the test selected.

In a strictly mathematical sense, the experimenter (E) maps a set of elements which are non-numerical but of conceptual interest in psychology onto a set of numbers when he measures a psychological variable. The mathematician would insist that E specify (1) the number system onto which he maps (such as integers or all rational numbers, etc.) and (2) what arithmetic operations and relations are implied in the mapping. This second point indicates whether order is present and whether the numbers can be added and/or multiplied meaningfully. Es who are "fixated" at this level commonly worry about nominal versus ordinal scales, etc., but their worries are rarely useful since little is known about the operations and relations that hold for the original variable.

[1] Now at University of Michigan.

Knowledge of the number system plus an assumption of how the probabilities of the numbers are related gives the mathematician a probability density function of the numerical variable—this is the basic mathematical model of the statistician. Given this model (e.g., binomial, Poisson, or normal) the statistician can then specify the probability of any admissible statistic describing a random sample taken from that population of numbers. *E*s "fixated" at this level worry about the distribution of their sample data—again with rare knowledge of what the population really looks like.

Part of the science of data analysis, then, is to choose the best model of the "real world" numbers which have resulted from empirical measurement. Obviously, however, there are only a finite number of such models in existence and there are many more ways in which the data may occur. If *E* always mapped psychological elements of his universe into the number "43" one model would suffice . . . a trivial one indeed. More commonly *E* maps elements into either integers or rational numbers, and the resulting empirical density distributions look "something like" the normal curve model.

This paper is concerned with the situation in which *E* administers different treatments to independent random samples drawn from the same population and wishes to test the hypothesis: $\mu_{X1} = \mu_{X2} \cdots = \mu_{Xk}$. This implies that *E*'s original interest in the investigation is formulated in terms of means of a chosen measure, X, (whether X be the original behavioral measure or some transformation of it), and that the alternative hypotheses of interest also are expressed in terms of the means on this same measure. This hypothesis may be tested by obtaining the mean square between groups divided by the mean square within groups (ms_b/ms_w) from a simple analysis of variance, and tables of the F distribution.[2]

When the empirical data or *a priori* analysis suggest a deviation from the normal curve model used in the mathematical derivation of this statistic, *E*'s have often chosen to apply a "normalizing

---

[2] If $k = 2$, this hypothesis could be tested by a two-tailed t test for independent groups. Squaring the obtained value of t would yield the value of ms_b/ms_w and the outcome would be identical to that of an F test. The use of the words "F test" in this paper may thus be considered as including this particular t test. The usage does not cover repeated measures designs in analysis of variance, or the dependent measures t.

transformation" rather than carry out the F test on the non-normal X's. In doing so, they are using a double mapping procedure which takes psychological elements to numbers on which the hypothesis is made (such as equality of means) and then takes the numbers into some function (such as logarithms) which is "closer to normality." In this double mapping procedure, the statistic will be derived from the second distribution, but the inference is made on the first distribution. This procedure might be defended if it displayed superiority to the F test on the original non-normal data on two criteria: (a) more exact control of the risk of a Type I error (rejecting a true null hypothesis). (b) Less risk of a Type II error (retaining a false null hypothesis), i.e., a higher power for a given degree of violation of the null hypothesis and a given level of significance. This study covers both criteria by contrasting the power curves of the F test as applied to the original non-normal data, and as applied to the "normalized data."

Curvilinear Transformations

The application of a non-linear function, $Y = f(X)$, will alter the variance to a degree relative to the mean of the population. The desired result of equal variances *and* normality is obtainable theoretically under only two conditions (Lindquist, 1953): (1) When the null hypothesis is true and all treatment populations have the same form and variance, and (2) When the null hypothesis is false and all treatment populations fit the same distribution function such that the variances can be shown to be a function of the means, $\sigma_j^2 = f(\mu_j)$. Given one of these situations, a continuous measure, and an exact knowledge of the distribution function of the populations, a mathematician can then determine the transformation that will reduce the data to a normal distribution. In rare cases this transformation will be of a simple form and perhaps even one of the commonly-used procedures ($Y = \log X$, $\sqrt{X}$, $1/X$); but since these are very specific instances, the expected transformation usually will be some complex function. A practical problem then is that E is usually unable to specify exactly the population condition, and, with the lack of information about the population, the proper transformation is usually indeterminate.

Considering the second Lindquist condition of correlated means and variances, a rough correction can be found if E is not interested

in the form of the transformed variables. Curtiss (1943) and Bartlett (1947) have given several mathematical rules which can be followed under these circumstances. In this case, if the treatment populations are distributed with different means and different forms (skewness and kurtosis) there is also some chance that the resulting transformed variables will be normal. However, if they are of homogeneous form, the different levels of the scores will be affected differently by the transformation, and the result will be heterogeneous forms on the transformed variate, only one of which might be normal. The converse is true in the other Lindquist condition where the null hypothesis is true. Here, if the forms of raw data are heterogeneous and there is no treatment effect, they will transform to heterogeneous distributions.

Once the transformation is made and the data analyzed, a further theoretical problem confronts *E* in the interpretation of the statistics obtained on the transformed scale. Since the statement that $\mu_{X1} = \mu_{Xk}$ is not equivalent to the statement that $\mu_{f(X1)} = \mu_{f(Xk)}$ in general, it is not logically correct to transfer conclusions from the statistical analysis of transformed measures back to the original hypothesis *E* started with. An example where the first statement is true and the second is not is given in Table 1.

TABLE 1

Relationships between Hypotheses for Two Populations of Data

Population		log Transformations to Base 2	
X_1	X_2	log X_1	log X_2
2	8	1	3
2	8	1	3
2	8	1	3
2	8	1	3
2	8	1	3
32	2	5	1
$\Sigma X_1 = 42$	$\Sigma X_2 = 42$	$\Sigma \log X_1 = 10$	$\Sigma \log X_2 = 16$
$\mu_{X_1} = 7$	$\mu_{X_2} = 7$	$\mu_{\log X_1} = 1.67$	$\mu_{\log X_2} = 2.67$
$\mu_{X_1} = \mu_{X_2} = 7$		$\mu_{\log X_1} \neq \mu_{\log X_2}$	

If *E* randomly and independently assigns subjects from the same parent population to his *k* treatments, and if the *k* treatment effects are additive and identical, then $\mu_{X1} = \mu_{X2} = \cdot \cdot \cdot = \mu_{Xk}$ and in addition, for any one-to-one transformation it also follows that $\mu_{f(X1)} = \mu_{f(X2)} = \cdot \cdot \cdot = \mu_{f(Xk)}$. This result occurs because the *X*

populations are identical to each other, and, similarly, the transformed populations are identical to each other. However, if treatment effects are not identical, or are not additive (that is, the effects of the treatments differ from subject to subject, certainly a common state in psychology), then the X distributions will no longer be identical. If this is the case, then the hypothesis of equal means on X may be true, while the hypothesis of equal means on $f(X)$ is false, and vice versa. Rejecting the hypothesis of equal treatment means on $f(X)$ may occur because the treatment populations on X differ in variance, or in skewness, or in kurtosis, even though the population means on X are equal. Thus rejecting the hypothesis that $\mu_{f(X1)} = \mu_{f(X2)} = \cdot \cdot \cdot = \mu_{f(X3)}$ implies some difference in the treatment effects, but does not clearly imply the rejection of $\mu_{X1} = \mu_{X2} = \cdot \cdot \cdot = \mu_{X3}$.

The E who is thus interested in the hypothesis of equal means on X, but who transforms his measure merely due to form considerations, thus ends up testing a different hypothesis than originally intended, and is typically not logically justified in extending his conclusions back to the original measure of interest. The material on transformations in many common statistical texts used by psychologists is relatively meager (Edwards, 1960; Guilford, 1956; Ray, 1962; Walker and Lev, 1953), and few cautions against this practice are found.

Method

The empirical sampling study which follows contrasts the power curves of the hypothesis $\mu_{X1} = \mu_{X2} = \mu_{X3}$ for various violations of the normality assumption, when tested by (a) the usual ms_b/ms_w on the X data or by (b) the ms_b/ms_w ratio on the $f(X)$ data, where $f(X)$ is the appropriate transformation that will "normalize" all of the k populations when the null hypothesis is true. Since our hypothesis is about parameters defined on the X scale, the points of the power curve are defined in terms of deviations of the means in units on the X scale. This is representative of the situation where the treatment effects, if any, are strictly additive on X, so that all treatment population variances will be homogeneous and equal to the variance of the parent population. In short, we are concerned with the situation where all the assumptions of the fixed treatments analysis of variance model are met except the assumption of con-

tinuous measures and the assumption that the common population form is normal. The discrete X measures are assumed to be evenly spaced. To decrease the possibility of the Central Limit Theorem obliterating any effects, the sample size was purposely kept small; $n =$ three or six.

The procedures used were similar to Boneau's (1962), except that a constructed population of 2560 measures was stored in 2560 memory locations of an LGP-30 computer. A random selection procedure was then used to draw memory locations and treat the value stored in that location as X. This procedure was based on a residue-class pseudo-random number generator with a period equal to the square of 5881. Care was taken to avoid starting the sampling at the same point in the generation sequence during the course of collecting the data. The numbers generated were reduced to 12-bit sequences, which were interpreted as random addresses of the X values to be selected.[3] The procedure is the equivalent of random sampling from a population of an infinite number of cases but with the exact characteristics (discrete measures and finite range) of the relative frequency distribution of the 2560 measures.

To each of the n measures in a given treatment group, the specified treatment effect (t_1, t_2, or t_3) was added, and the ms_b/ms_w ratio was computed on the resulting measures. This process was repeated until one thousand ratios were computed. The computer output was a frequency distribution of mean square ratios accompanied by an exact count of the ratios exceeding critical F values in common usage: $F_{.05}$, $F_{.025}$, $F_{.01}$, $F_{.005}$ for $k - 1$, $k(n - 1)$ degrees of freedom.

For simplicity, the treatment effects were specified as $t_1 = 0$, $t_2 = +a$, and $t_3 = +2a$; thus, a is the distance between the population means in units on the X scale. By varying the arbitrary constant, a, it is possible to study any desired points on the power curve. The degree of violation of the null hypothesis is conventionally expressed by the value ϕ, defined as

$$\phi = \sqrt{\frac{n \sum_{i} (\mu_X + t_i - (\mu_X + \bar{t}))^2}{k\sigma_X^2}}$$

where $\bar{t}$ is the average of the t_i's.

[3] After this study was completed, it was found that the method of extracting random addresses led to numbers not uniformly distributed across the 64 sec-

Given the value of ϕ and treatment effects specified in a linear fashion as above, it is possible to solve for the needed values of a by the equation:

$$a = \sigma_X \phi \sqrt{\frac{12}{n(k^2 - 1)}}.$$

Thus, when ϕ is zero, a is zero, and the parameter condition $\mu_X + o = \mu_X + a = \mu_X + 2a$ is that of a true null hypothesis. The observed proportion of "significant" F's obtained when $\phi = 0$ is thus an estimate of the actual risk of a Type I error. When ϕ is greater than 0, the null hypothesis is false, and the larger the observed proportion of "significant" F's, the lower the risk of a Type II error. Five values of ϕ were studied, 0, .5, 1.0, 1.5, and 2.5. The increasing values of ϕ are proportionate to the increasing distance between the population means (a).

Six parent populations of varying forms were used to determine the power curves of the F test under various violations of normality. One population was a discrete approximation to the normal curve, and primarily served to check the procedures used and any effects due to discreteness. Symmetrical rectangular and leptokurtic populations were used to determine the effects of kurtosis, and three unimodal populations varying in skewness (and accompanying leptokurtosis) were used. These three skewed populations were constructed in such a manner that, when the treatment effect was zero, they could be transformed to near normality by three commonly used procedures: log, square root, and reciprocal. These distributions are labeled as slight skew ($\sqrt{\ }$-N, meaning that the square root transformation would "normalize" the distribution), moderate skew (log-N), and extreme skew (reciprocal-N). It was thus possible to make use of transformations in a more satisfactory manner than the typical E, in that the transformation used was definitely appropriate for the common population shape when the null hypothesis was true. Of course, when the null hypothesis is false, the transformation was an appropriate "normalizer" only for the first treatment population, but this is a hazard implicit in the use of transformations in this situation. Table 2 shows the de-

tors or word spaces of the memory tracks. When this was corrected, the new power functions constructed for the normal and leptokurtic populations were practically identical to the original. It was concluded that the effect on this study due to the biased procedure was negligible.

scriptive parameters of the six parent populations, with the parameters of the "transformed populations" listed beneath their appropriate skewed populations. The means and variances of these distributions are irrelevant since they have nothing to do with the assumption about form except in determining the effects of the transformations, and the definition of ϕ assesses differences in variance to establish a similar degree of violation of the null hypothesis for all populations. To study the effects of analyzing transformations of the data on the probability of rejecting a null hypothesis based on the raw data (X), the treatment effects were added prior to the transformation. Thus, these computations were carried out on $Y = f\ (X \times t_j)$.

Skewness is reflected in the parameter, β_1. A symmetrical distribution will yield zero values of β_1, while positive values indicate a concentration of scores to the left of the mean and/or a higher range of scores above the mode than below it (positively skewed); negative values indicate the opposite condition. The kurtosis parameter (β_2) is a measure of degree of concentration of scores in the distribution. In the normal curve, $\beta_2 = 3.0$. Values above this are characteristic of "leptokurtic" distributions—ones highly concentrated at a single mode and having long tails. Values of $\beta_2 < 3$ indicate "platykurtic" distributions—ones approaching uniform or multimodel shapes. The mathematical specifications of β_1 and β_2 follow Pearson (1931). Although it is very common to obtain leptokurtic or platykurtic distributions with no skewness, most unimodal skewed distributions also show leptokurtosis. Thus on empirical data, skewness is typically accompanied by leptokurtosis.

It may be noted in Table 2 that the discrete normal approximation and those distributions transformed to normality show a kurtosis index slightly below the expected 3.00. This is apparently due to the finite size of the populations, which necessitated the exclusion of extreme values with a probability of less than 1/2560, and thus truncated the ends of the distributions. The discrepancy between the continuous distributions of mathematical theory and the discrete distributions of recorded data is brought out another way when considering transformations. Figure 1 shows the discrete extremely skewed (reciprocal-normal) distribution, contrasted to the "normalized" transformed distribution. The "normalizing" of such discrete distributions can only be accomplished by spreading out

points at one portion of a scale and contracting points at other portions. In this example, there are only eight score values covering the distance to the left of the mean of the "normalized" distribution, while 52 values are found to the right of the mean. The percentile rank of the mode of this "normal" distribution is 36.3.

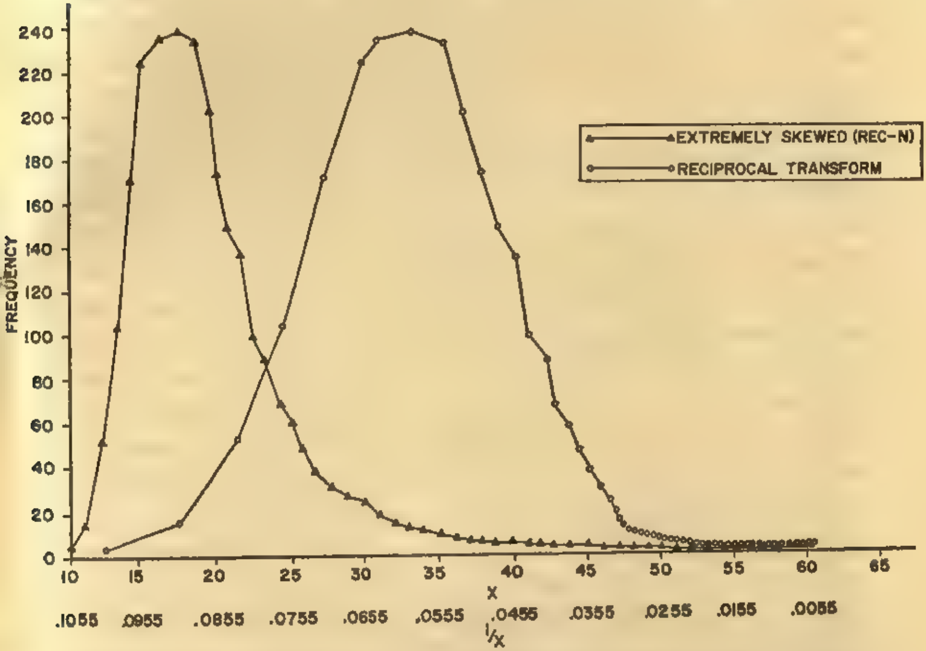


Figure 1. Frequency polygons for the extremely skewed (Rec-N) population and its corresponding "normalized" population, the reciprocal-transform.

It should be noted that $f(\mu_X) = \mu_{f(X)}$ only when the X population meets very restricted conditions specific to the transformation involved. For example, for the case $Y = \log X$: $\mu_Y = \log \mu_X$ only if $\mu_X^n = \pi^n X$, the product of all the X values. Thus, in general, E can not take the complementary transformation of the mean of the transformed variable as equivalent to μ_X. For the distribution in Figure 1, the reciprocal of the mean of the $1/X$ distribution is 18.5, about 1.4 points below the population mean of the X distribution.

Results

All ms_b/ms_w distribution obtained when the null was true were tested for goodness of fit to the theoretical F curve with appropriate degrees of freedom. Only two of the 18 tests were significant at the

.05 level, those for samples of size $n =$ three from the leptokurtic population and the rectangular population (Chi Squares of 38.91 and 65.08 respectively, both with df $= 24$). However, since these tests are not sensitive to deviations in the tails of the sampling distributions, the area of greatest interest, the results are best expressed in terms of the proportions of the sampling distributions that exceeded the theoretical critical values of F, or the empirical estimate of the power of the test under the given condition. Tables 2 and 3 show the empirical power estimates using the .05 significance level with an n of three and n of six respectively.[4] Theoretical power values were approximated for the non-central F distribution using a procedure developed by Severo and Zelen (1960). Critical values for the significance of deviations in power were computed for the .05 level and are listed in the bottom row of each table. Deviations in empirical estimates from the theoretical power are listed and are marked with an asterisk when significant.

It is evident from Tables 2 and 3 that the normal distribution sampled in this study resulted in power values not significantly different from the theoretical. This indicates that the procedures used in this study were sufficiently valid to make consistent deviations meaningful indications of the effects of non-normality. It can also be seen that the slightly skewed and moderately skewed populations gave very little deviation from theoretical power for both sample sizes, with a general trend to increase the power in the range of moderate degrees of falsity of the null. Transformations of these two distributions show slightly greater effect, usually in the same direction as that of the original distribution.

The greatest deviation from theoretical power curves resulted from sampling the two highly leptokurtic populations, the extremely skewed, where $\beta_2 = 9.54$, and the symmetrical leptokurtic, where $\beta_2 = 9.16$. All four curves involving these populations show the test to be slightly conservative when the null is true ($\phi = 0$), but to be appreciably higher in power than expected when in the re-

[4] Similar tables were obtained for the 2.5%, 1%, and .5% significance levels; however, the lower values of the expected number of ratios exceeding the tabled values makes the trends found in these tables less stable than those shown. Complete frequency distributions of all populations used, and the 2.5%, 1%, and .5% tables have been deposited with the American Documentation Institute. Order Document No. 8790, remitting \$1.25 for 35-mm. microfilm or \$1.25 for 6 by 8 in. photocopies.

TABLE 2

Proportions of the Obtained ms_b/ms_w Distributions Greater Than the Tabled 5% F Value (5.14) $k = 3, n = 3$

POPULATION	PARAMETERS				φ VALUES									
					0		.5		1.0		1.5		2.5	
	μ	σ	β_1	β_2	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.
Theoretical					.050		.087		.206		.417		.851	
Normal	31.00	9.92	0.00	2.90	.046	−.004	.098	.013	.200	.006	.430	.013	.851	.000
Leptokurtic	50.00	7.28	.00	9.16	.049	−.001	.101	.014	.269	.063*	.522	.105*	.874	.023*
Rectangular	31.50	18.19	0.00	1.80	.070	.020*	.106	.019*	.209	.003	.400	−.017	.856	.005
Slt. Skew (√ -N)	39.50	12.38	.45	3.53	.052	.002	.077	−.010	.193	−.013	.410	−.007	.848	−.003
√ -Transform	6.20	.99	0.00	2.91	.054	.004	.086	−.001	.201	−.005	.518	.001	.834	−.017
Mod. Skew (Log-N)	23.48	5.31	.64	3.53	.058	.008	.080	−.007	.188	.018	.446	.029	.858	.007
Log-Transform	3.13	.22	0.00	2.82	.058	.008	.094	.007	.213	.007	.445	.028	.828	−.030*
Ext. Skew (Rec-N)	19.98	6.36	2.04	9.54	.048	−.002	.108	.021*	.264	.058*	.541	.124*	.854	.003
Rec-Transform	.054	.014	.03	2.88	.059	.009	.134	.047*	.325	.119*	.536	.119*	.792	−.059*
Critical Deviation ($p < .05$):						.0135		.0175		.0251		.0356		.0221

* Signficant at .05 level

TABLE 3

Proportions of the Obtained ms_b/ms_w Distributions Greater Than the Tabled 5% F Value (3.68) $k = 3$, $n = 6$

POPULATION	ϕ VALUES									
	0		.5		1.0		1.5		2.5	
	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.	Pwr.	Dev.
Theoretical	.050		.097		.261		.540		.954	
Normal	.059	.009	.091	−.006	.275	.014	.519	−.021	.945	−.009
Leptokurtic	.036	−.014*	.109	.012	.338	.077*	.584	.044*	.924	−.030*
Rectangular	.058	.008	.094	−.003	.258	−.003	.538	−.002	.954	.000
Slt. Skew ($\surd$-N)	.052	.002	.096	−.001	.266	.005	.556	.016	.947	−.007
$\surd$-Transform	.046	−.004	.098	.001	.286	.025	.575	.035*	.946	−.008
Mod. Skew (Log-N)	.054	.004	.100	.003	.273	.012	.563	.023	.943	−.011
Log-Transform	.057	.007	.097	.000	.268	.007	.594	.054*	.941	−.013*
Ext. Skew (Rec-N)	.037	−.013	.078	−.019*	.326	.061*	.559	.019	.910	−.044*
Rec-Transform	.057	.007	.120	.023*	.485	.224*	.657	.117*	.964	.010
Critical Deviation ($p < .05$):		.0135		.0183		.0272		.0309		.0130

* Significant at .05 level

gion of moderate departures from the null ($\phi = 1.0, 1.5$). The power curves for samples of $n = 3$ from these populations are contrasted with the theoretical curve, and the curve for the platykurtic rectangular population in Figure 2. The rectangular, or platykurtic, distribution ($\beta_2 = 1.80$) shows an effect opposite to that of the leptokurtic cases, although it lies closer to the theoretical curve. The reciprocal transformation of the extremely skewed population (dashed line in Figure 2) in general surpassed the power of the test on the raw data, except at $\phi = 2.5$, where a loss is quite apparent. The increase in risk of a Type I error when $\phi = 0$ is generally true for transformations (causing an average increase of .006 over the risk on the raw data), although none of these differences is significant. Only in the case of the large sample size and extremely skewed distribution did a transformation improve the accuracy of the test when the null was true.

Discussion

This empirical sampling study was designed to help the experimenter interested in the equality of treatment effects on the X scale to choose between (1) conducting the usual ms_b/ms_w test on data known to be non-normal, or (2) transform the data by some procedure to a shape closer to normality and test the means of the transformed variate. Although the results obtained are specific to the particular populations used and undoubtedly contain some sampling variability, the results show consistent trends.

For the slightly skewed and moderately skewed distributions, the deviations from the theoretical power curve were so small that the extra trouble of using transformations would seem sufficient cause to abandon them. The fact that the use of the appropriate transformation increased the deviation from the theoretical power more often than it decreased it raises further questions about this practice. In an actual experiment, E typically has no adequate basis for choosing the "most appropriate" transformation. Furthermore, since this practice involves either the use of the unlikely assumption that the hypothesis is the same on X as on the transformed measures or else requires an embarrassing change of hypotheses at the end of the investigation, the authors see no justification for continuing this practice on such distributions.

The only substantial deviation from the normal power curve was for the two populations that were highly leptokurtic—the extremely skewed population and the symmetrical leptokurtic population. The power curves for tests on these populations are consistent with the theoretical results of Srivastava (1959) for continuous distributions: the risk of a Type I error is slightly less than specified, with the negative deviation at $\phi = 0$ changing to positive deviations at points of intermediate power, and with a second crossing of the theoretical power curve occurring at high power points when the deviation again becomes negative (as at $\phi = 2.5$ when $n = 6$).

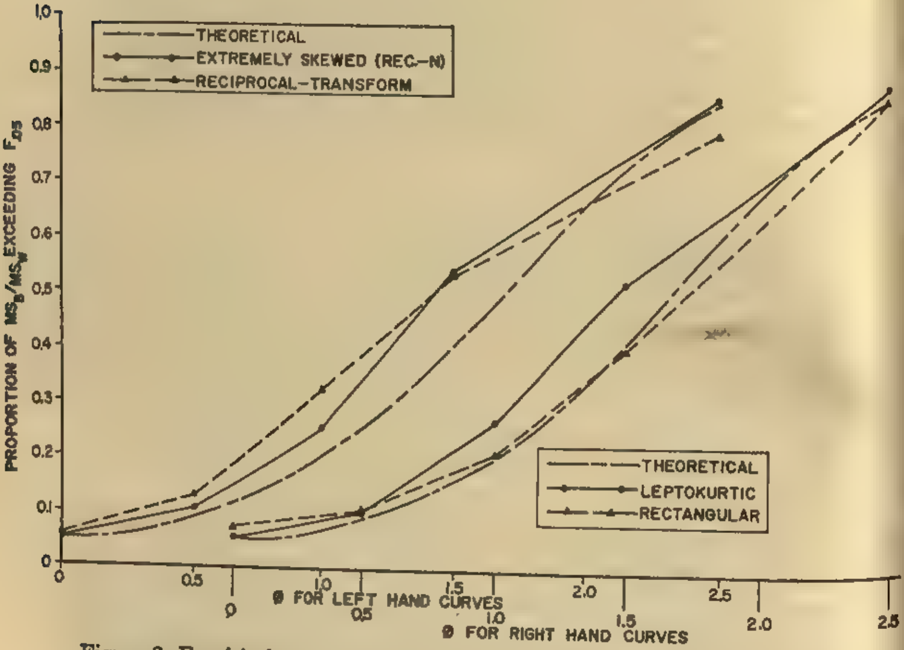


Figure 2. Empirical power curves for certain nonnormal populations, $k = 3$, $n = 3$.

Thus, the trends found by Srivastava may be extended to discrete populations with considerably higher β_2 values than could be investigated by his methods. Oddly enough, these results suggest that leptokurtic data distributions might improve the F test. The conservative test when the null is true would be to err on the safe side, and the increased power through moderate degrees of falsity of the null would give the researcher greater power where it is typically

most needed (Cohen, 1962). The fact that the empirical power is slightly lower for very high power points is a slight disadvantage but one that *E* might be willing to exchange for relatively greater increases in power at intermediate levels. No simple common transformation could "normalize" a symmetrical leptokurtic distribution; the "normalizing" reciprocal transformation increased the amount of discrepancy from the theoretical curve in seven of ten comparisons.

The problem of platykurtosis may be judged more serious by many *E*s, since this condition tends to have effects opposite to those above and thus increases the risk of a Type I error above the specified significance level. However, the authors would consider the deviations shown for the rectangular population on samples of six cases to be quite minor, and Srivastava's work (1959) indicates that increasing n decreases the kurtosis effects on the power curve. The authors feel the rectangular population ($\beta_2 = 1.8$) is quite extreme for psychological data and would feel reasonably safe using the conventional F test with equal n's of ten or more. If bimodal populations of extreme platykurtosis were encountered (minimum possible β_2 is 1.0), larger n's would be advisable.

The results of this study and of Srivastava's theoretical work contradict the Norton (1952) results showing that leptokurtosis increased the risk of a Type I error. The difference may be due to the fact that Norton sampled exhaustively, without replacement, from a finite population, while the present study sampled randomly, with replacement, from a finite population.

It should be emphasized that the present study has dealt only with use of transformations to correct for violations from normality. There are many other usages of transformations that have not been, and are not being, criticized. The use of a transformation is essentially selecting a scale of measurement. Mueller (1949) pointed out that some theories may specify the use of a transformation by, say, using log X as the scale of measurement in the theory. To test the theory then, measures should be transformed as required. Other times, a transformation may be used to simplify the expression of a relationship that requires a more complex equation when expressed in X. Occasionally, a curvilinear relation between W and X may be linear between W and log X; if the homoscedasticity also improves, then a considerable simplification in

description is possible. These are pertinent reasons for selection of a scale of measurement; what is being said is that often the fact of non-normality alone is NOT a sufficient reason for changing from an *a priori* acceptable scale of measurement to some other arbitrary scale. In the analysis of variance of independent groups, in particular, the disadvantages of this procedure seem to outweigh the advantages.

In addition to normality, transformations have also been recommended to achieve homogeneous variances, or additivity of treatment effects. It should be noted that these three goals will not necessarily be met by the same transformation, although it is possible for this to occur. Box (1954) has shown that homogeneous variances are not too important in analyses of variances of independent groups with equal n's, although this goal may be more important in repeated measures designs and prediction uses involving homoscedasticity. Similarly, lack of additivity does not disable a simple analysis of variance of independent groups. However, additivity would be quite important in a design that assumed no interactions.

In general, the use of a clearly interpretable scale of measurement certainly should be the dominant consideration. If the statistical characteristics of the scale are not ideal, often an experimental design that is not greatly impaired by the defect may be found. Statements by Tippett (1952, p. 344–345) best summarize the argument. "If a transformed variate (Y), having convenient statistical properties can be substituted for X *in the technical arguments from the results and in their applications* (italics ours) there is everything to be said for making the transformation. If the technical interpretation has to be in terms of the untransformed variate X, and after the statistical analysis has been performed on (Y), means and so on have to be converted back to X, statistical difficulties arise, and the waters deepen. Readers are advised not to make transformations on statistical grounds alone unless they are good swimmers and have experience of the currents."

REFERENCES

Bartlett, M. S. The Use of Transformations. *Biometrics Bulletin*, 1947, 3, 39–52.

Boneau, C. A. A Comparison of the Power of the *U* and *t* Tests. *Psychological Review*, 1962, 69, 246–256.
Box, G. E. P. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: I. Effect of Inequality of Variance in the One-Way Classification. *Annals of Mathematical Statistics*, 1954, 25, 290–302.
Cohen, J. The Statistical Power of Abnormal-Social Psychological Research: A Review. *Journal of Abnormal and Social Psychology*, 1962, 65, 145–153.
Curtiss, J. H. On Transformations Used in the Analysis of Variance. *Annals of Mathematical Statistics*, 1943, 14, 107–122.
Edwards, A. L. *Experimental Design in Psychological Research.* (2nd ed.) New York: Rinehart, 1960.
Guilford, J. P. *Fundamental Statistics in Psychology and Education.* (3rd ed.) New York: McGraw-Hill, 1956.
Lindquist, E. F. *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton Mifflin, 1953.
Mueller, C. G. Numerical Transformations in the Analysis of Experimental Data. *Psychological Bulletin*, 1949, 46, 198–223.
Norton, D. W. An Empirical Investigation of the Effects of Nonnormality and Heterogeneity upon the *F*-test of Analysis of Variance. Unpublished doctoral dissertation, State University of Iowa, 1952.
Pearson, E. S. The Analysis of Variance in Cases of Non-normal Variation. *Biometrika*, 1931, 23, 114–133.
Ray, W. S. *Statistics in Psychological Research.* New York: Macmillan, 1962.
Severo, N. C. and Zelen, M. Normal Approximations to the Chi Squared and Non-Central *F* Probability Functions. *Biometrika*, 1960, 47, 411–416.
Srivastava, A. B. L. Effect of Nonnormality on the Power of the Analysis of Variance Test. *Biometrika*, 1959, 46, 114–122.
Tippett, L. H. C. *The Methods of Statistics.* (4th ed.) New York: John Wiley and Sons, 1952.
Tukey, J. W. The Future of Data Analysis. *Annals of Mathematical Statistics*, 1962, 33, 1–67.
Walker, H. M. and Lev, J. *Statistical Inference.* New York: Henry Holt, 1953.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

SOCIAL DESIRABILITY ESTIMATION AND "FAKING GOOD" WELL[1]

JERRY S. WIGGINS
University of Illinois

The apparently widespread tendency of subjects to put their best foot forward when responding to items from personality and adjustment inventories has remained largely unanalyzed in the wake of recent efforts to detect, discourage or bemoan such behavior (Hanley, 1957; Voas, 1958; Edwards, 1957). Variations in the apparent strength of this tendency under different circumstances of assessment suggest that the phenomenon is in part a situational one. Even within a single assessment context, however, considerable individual differences in this tendency may be observed, suggesting, among other things, a person-situation interaction which reflects motivational characteristics of the subjects. Moreover, there is some evidence which suggests that when subjects are explicitly told to make a good impression, those who do "best" are likely to be more psychologically sound than those who do less well at this task (Grayson and Olinger, 1957; Canter, 1963). The recent revival of interest in "role-playing" studies has produced a considerable amount of data which supports the observation that even under instructions designed to induce a uniformly high degree of good impression motivation, subjects will vary extensively in their performance (Wiggins, 1959; Walker, 1961; 1962). The present study was designed to test a hypothesis concerning one possible source of these observed individual differences.

Whether "psyching out" a potential employer or cooperating with the explicit instructions of a role-playing experiment, making a

[1] This investigation was supported by Public Health Service Research Grant No. MH 07042-01 from the National Institute of Mental Health.

good impression via a personality inventory would seem to involve an ability or "shrewdness" component as well as a motivational one. Given the motivation to fake, one must know the standards of the test scorer with respect to what will be judged favorably. The social desirability rating procedures developed by Edwards (1953) provide an excellent paradigm for discussing this process. In Edwards' terms, the subject must attempt to *estimate* the social desirability value of an inventory item and respond accordingly (Edwards, 1957). A correlated, and perhaps indistinguishable, hypothesis would be that the subject attempts to estimate the proportion of people who would admit to the item (communality value) and responds accordingly (Wiggins, 1962). Whether the subject is estimating "desirability" or "popularity," he is nonetheless estimating and he may do so accurately or inaccurately. The present study tests the hypothesis that *accuracy* of desirability and communality estimations are positively related to the *degree* of successful faking a subject attains under explicit instructions to make a good impression.

Method

Subjects

The subjects were 239 male undergraduates who were assigned to this study early in the semester as part of the course requirement in introductory psychology at the University of Illinois. Subjects were randomly assigned to one of six experimental conditions which each required approximately one hour of their participation.

Experimental Design

The experimental conditions differed both in the tasks required of the subjects and in the order in which these tasks were presented. Two sets of stimuli (items) were presented to all but one of the six groups thus formed. One set of stimuli consisted of 50 items and will be referred to as the *rating form.* The other set of stimuli consisted of 212 items and will be designated the *inventory*. Three sets of instructions were employed with the rating form: (a) estimate communality, (b) estimate social desirability or (c) rate social desirability. Two sets of instructions accompanied the inventory: (a) self-report or (b) fake good.

Table 1 illustrates the order of stimulus presentation and kinds of instruction which define the six separate subject groups. The Communality Estimation group estimated communality on the rating form and faked good on the inventory. Half of the subjects in

TABLE 1

Experimental Design ($N = 239$)

GROUP	DESIGNATION	N	ORDER OF PROCEDURES	
Communality Estimation	CE_1	21	Est. Comm.	Fake
	CE_2	25	Fake	Est. Comm.
Desirability Estimation	DE_1	23	Est. SD	Fake
	DE_2	25	Fake	Est. SD
Desirability Rating	DR	50	Rate SD	Self-Report
Self-Report	SR	95	Self-Report only	

this group estimated communality first (CE_1) and the other half faked good first (CE_2). The Desirability Estimation group estimated desirability on the rating form and faked good on the inventory. The order of stimulus presentation was also counterbalanced within this group. Subjects in the Desirability Rating group first estimated desirability and then gave their self-report to the inventory. The Self-Report group simply gave their self-report to the inventory.

Rating Form

The 50 items employed for the rating form did not overlap with those appearing in the inventory. The rating form items were selected in such a way as to maximize covariation in rated social desirability value and endorsement percentage. There were three versions of the rating form which were alike in that the same items were to be judged on an 11-point scale. The versions differed in instructions and in the descriptive anchors accompanying the 11-point scale. The three rating tasks were as follows:

Communality Estimation (CE). Subjects who received this version of the rating form were asked to give their best estimate of the *proportion* (per cent) of men in their class who would answer True to each of the 50 statements. They did so on an 11-point rat-

ing scale of percentages: 0%, 10%, 20% . . . 100%. These subjects were, in effect, being asked to estimate the endorsement percentages which would occur in a group such as the SR group.

Desirability Estimation (DE). Subjects who received this version of the rating form were asked to estimate which of a number of rating categories would best represent the opinion of men in their class with respect to the *desirability* of a True answer to each of the 50 statements. Ten rating categories were provided in the form of an 11-point scale: 0 (very extremely undesirable), 1 (extremely undesirable), 2 (strongly undesirable) . . . 10 (very extremely desirable).

Desirability Rating (DR). Subjects who received this version of the rating form were asked to give *their* opinion of the desirability of a man in their class answering True to each of the 50 statements. The same 11-point rating scale of desirability employed by the Desirability Estimation (DE) group was provided. In a sense which will be amplified later, the subjects in the DE group were trying to estimate the *pooled* or averaged values obtained from the DR group.

Inventory

The 212 item inventory was comprised of items from MMPI scales which had previously been shown to be sensitive to role-playing instructions. Rosen (1956a and b), Walker (1961; 1962) and Wiggins (1959) have presented data on MMPI clinical and stylistic scales which exhibit significant shifts in scale means under role-playing instructions. The first two of these studies obtained repeated measurements (standard-fake) on the same subjects while the third study contrasted the scale scores of independent groups (standard and fake). Scales were selected for the present inventory from among those which had exhibited significant mean score shifts under role-playing instructions in at least two of the three independent studies (Rosen, 1956a and b; Walker, 1961; 1962; Wiggins, 1959). MMPI clinical scales which exhibited a significant decrease in these studies were *Pt* (McKinley and Hathaway, 1942), *A* (Welsh, 1956), *Hy-O* and *D-O* (Wiener, 1948). MMPI clinical scales which significantly increased were *Hy-S* and *D-S* (Wiener, 1948). MMPI stylistic scales which increased significantly under role-playing instructions were *Sd* (Wiggins, 1959),

Mp (Cofer, Chance and Judson, 1949) and *L* (Meehl and Hathaway, 1946).

The 212 item pool which included all items from the preceding nine scales was presented in randomized order in the standard inventory format with two response options (True and False). Two sets of instructions were employed with the True-False inventory:

Self-Report (SR). The self-report instructions were adapted from those utilized in the standard administration of the MMPI (Hathaway and McKinley, 1951). Subjects were instructed to mark a statement True if it applied to themselves and False if it did not.

Faking Instructions. The role-playing instructions employed in the present study attempted to maximize the amount of social desirability responding in accord with suggestions along this line made by Walker (1962). The instructions read, in part:

> "We want to see how well you can do in *answering these statements so as to create the most favorable impression you can.* Pretend that your answers to these statements will be read out loud in class. Assume also that you are *especially concerned* with making a good impression on your classmates. Now answer *each statement* in such a way as to create the best impression on others in the class . . . Remember, you are *NOT* being asked whether the statements are True or False as applied to yourself. Rather, you are being asked to answer each statement in such a way as to *create the best possible impression* on others in your class."

Accuracy Measures

Communality Estimation. As previously mentioned, 95 subjects in the group designated SR took an inventory under self-report instructions only. This inventory was an augmented version of the 212 item inventory which included, in addition, the 50 items from the rating form; here in true-false format. Item endorsement percentages were computed for these 50 items in the group of 95 standardly instructed subjects. These endorsement percentages were rounded to the nearest tenth (0%, 10%, 20% . . . 100%) to serve as the criterion indices of accuracy of communality estimation. An absolute accuracy score (AA) was computed for each subject by summing the discrepancies between estimated and obtained endorse-

ment percentage across the 50 items without regard for the direction of such discrepancies. In addition, a relative accuracy score (RA) was computed across all 50 items which was based on the algebraic differences between estimated and obtained endorsements.

Desirability Estimation. As already indicated, 50 subjects in the DR group were asked to give their own opinions as to the desirability (11-point scale) of their classmates subscribing to the 50 rating form statements. The median desirability values of the subjects in this group served as the criterion indices of accuracy of desirability estimation for subjects in the DE group. These median desirability ratings were rounded to units (0, 1, 2, . . . 10) for purposes of comparison with the individually estimated desirability ratings of the DE group. Two indices of desirability estimation accuracy were computed in the same manner as those for communality estimation accuracy. These were: an absolute accuracy score (AA) based on summed absolute discrepancies between estimated and obtained desirability values, and a relative accuracy score (RA) based on summed algebraic discrepancies.

Faking Measures

A total of 94 subjects was given instructions to make a good impression on the 212 item inventory in a design counterbalanced for the possible order effects of the additional tasks of desirability and communality estimation. The somewhat knotty problem of operationalizing faking "success" was met by resort to scale scores on indices which had been previously demonstrated to be sensitive to role-playing instructions in both test-retest and contrasted group studies. High scores on the individual dissimulation scales *Sd, Mp* and *L* were considered evidence of faking success as was the sum of such "fake" scales: $Sd + Mp + L$. Low scores on the pathology scales *Pt, A* and *Hy-O* were taken as indicative of successful faking as was their sum: $Pt + A + Hy\text{-}O$. A gross index of faking success was also based on the difference between the preceding sum scores. Additional clinical scales were *Hy-S* and *D-S* which were expected to increase and *D-O* which was expected to decrease with successful faking.

In deference to the lack of unanimity of opinion as to the single "best" index of dissimulation (Wiggins, 1959; Hanley, 1961; Walker, 1962) several such indices were scored in an admittedly ex-

ploratory fashion. It should be noted, however, that the scale scores from the group of 95 subjects who took the inventory under standard instructions (SR) provided a baseline against which to evaluate the relevance of individual scales to the present study.

Results

Relevance of Faking Indices

The nine indices of faking success which had been selected from previous studies were scored separately for the present sample of 95 standardly instructed men and the 94 men who had received faking instructions. Table 2 presents mean scale scores for the two

TABLE 2

Mean Scale Scores on Faking Indices in Fake Good and Control Groups

Group		*L*	*Mp*	*Sd*	*Pt*	*A*	*Hy-O*	*Hy-S*	*D-O*	*D-S*
Fake Good	Mean	7.46	21.42	24.45	6.46	5.62	2.28	17.88	5.68	11.65
($N = 94$)	σ	3.51	5.45	6.09	6.05	5.46	3.96	3.51	4.48	2.21
Control	Mean	2.87	11.85	12.81	15.02	14.54	6.25	14.97	9.44	10.19
($N = 95$)	σ	1.84	3.95	3.73	6.90	7.76	3.38	3.95	4.45	2.88
	t	11.20	13.87	15.73	9.01	9.10	7.35	5.39	5.78	3.95
	p	.001	.001	.001	.001	.001	.001	.001	.001	.001

groups along with the t values for the differences between the groups. The hypothesis that the obtained mean scores for the faking and control groups could have arisen from the same population was rejected at less than the .001 level for all nine faking indices. This expected result is necessary but not sufficient to justify the employment of these measures as indices of faking success in the present study.

Other Effects

The counterbalanced nature of the present design permitted an assessment of the influence of the tasks of communality and desirability estimation upon subsequent faking success as measured by the several dissimulation indices of the inventory. Mean scale scores of subjects who had participated in the estimation procedures prior to the faking situation (Groups CE_1 and DE_1) were contrasted with mean scale scores of subjects who faked the inventory prior to the estimation procedures (CE_2 and DE_2).

For subjects in the Communality Estimation groups, the mean faking scale scores were highly similar for the group which faked prior to estimation and the group which faked after estimation. None of the nine mean scale score differences approached significance.

In contrast, three of the nine dissimulation scales were found to have different means under different orders of administration in the Desirability Estimation groups. Subjects who had desirability estimation experience *prior* to faking exhibited *higher* mean dissimulation scale scores on *Sd* ($t = 4.22, p < .001$), *Mp* ($t = 2.87, p < .01$) and *L* ($t = 1.80, p < .10$). None of the remaining dissimulation scale score differences approached significance.

Scales *Sd, Mp* and *L* have similar and interesting empirical properties which may be related to the present order effects. In role-playing studies they exhibit not only large mean score differences between faking and control groups, but high indices of discriminative efficiency as well (Wiggins, 1959). Further, they define a distinct factor of the MMPI which is generally agreed to be a dissimulation factor (Edwards, Diers and Walker, 1962; Wiggins, 1964; Liberty, Lunneborg and Atkinson, 1964).

A common property of *Sd, Mp,* and *L* as well as other dissimulation scales belonging to this factorial grouping (such as the Marlowe-Crowne (1960) Social Desirability Scale) is that item desirability values tend to exceed empirical endorsement frequencies. This fact has been variously emphasized as reflecting the "*obviousness*" of such items (Hanley, 1961), the *dishonesty* of subjects answering such items (Edwards et al, 1962) and the psychometric properties of items which can in fact *change* under special instructions (Messick and Jackson, 1961; Wiggins, 1962). The present results indicate that specific practice in *social desirability* estimation will enable subjects to achieve higher scores on faking scales subsequently administered. Such results do not allow a choice among the alternate emphases given to these dissimulation scales.

Accuracy and Faking

Social Desirability Estimation. The substantial and unanticipated differences in dissimulation scale scores between the counterbalanced Desirability Estimation groups required separate analyses of the relation between accuracy and faking in these two groups. Correlations between the two accuracy measures and the several indices of

TABLE 3

Correlations between Two Measures of Accuracy of Desirability Estimation and MMPI Indices of Faking Success for Subjects who Estimated Prior to Faking (DE_1) and Subjects who Faked Prior to Estimating (DE_2)

		L	*Mp*	*Sd*	ΣF	*Pt*	*A*	*Hy-O*	ΣP	ΣF-ΣP	*Hy-S*	*D-O*	*D-S*
Group DE_1	AA	−04	−10	−25	−16	47*	48*	29	46*	−36	−48*	22	−25
(*N* = 23)	RA	−35	−39	−30	−38	41	36	12	34	−44*	−23	19	−18
Group DE_2	AA	−14	−15	−18	−17	15	17	−14	10	−16	07	−07	−20
(*N* = 25)	RA	−21	−18	−11	−17	04	19	−01	11	−16	−10	−05	−10

* $p < .05$

faking success are presented in Table 3 for the group which estimated desirability prior to faking (DE_1) and the group which estimated desirability after faking (DE_2). In the group which estimated prior to faking, relative accuracy (RA) is related consistently to the dissimulation scales (*L*, *Mp*, *Sd*).[2] This suggests that tendencies to overestimate the social desirability of items are reflected in increased dissimulation scale scores under faking instructions. This trend is *not* statistically reliable in this small group of subjects, however.

In contrast with the findings for dissimulation scales, several of the pathology indices are consistently and reliably related to accuracy of social desirability estimation in Group DE_1. Absolute accuracy scores (AA) are positively and reliably correlated with reductions in *Pt*, *A* and Sum Pathology and increases in *Hy*-subtle under faking instructions. The same trend is apparent for the relative accuracy measure. When *S*s are asked to estimate the social desirability of a group of items prior to faking good on a different inventory, their accuracy of such estimation is related to their subsequent success in obtaining low pathology scores (*A*, *Pt*) and high scores on a subtle hysteria scale (*Hy-S*).

A quite different picture emerges when the accuracy versus faking success correlation matrix is examined for the group which faked prior to estimating desirability (DE_2). Although some of the same patterns are present there are many reversals of trend and *none* of

[2] In interpreting the *signs* of these correlations it should be noted that the accuracy scores are actually "inaccuracy scores" so that negative relations are expected with dissimulation scales and positive relations with pathology scales.

the correlations attains statistical significance. When *S*s are asked to estimate the social desirability of a group of items *after* faking good on a different inventory, there is no consistent relationship between their accuracy of estimation and their prior success in obtaining low pathology scores or high dissimulation indices.

Communality Estimation. Since no order effect was detectable in the communality estimation group it was possible to base the correlations between accuracy and faking success on the combined group of 46 *S*s. The accuracy indices were slightly and often inconsistently related to the dissimulation indices. Relations between accuracy and pathology score reduction were consistent in sign but rather clearly not significantly different from zero. Only the relation between communality overestimation (RA) and increase in subtle hysteria (*Hy-S*) was significantly different from zero ($r = .36$). The accuracy with which *S*s estimate communality is not obviously related to their success in reducing pathology scores or increasing dissimulation indices when faking a separate inventory.

Discussion

The present findings do not permit a simple formulation of the relationships between accuracy of popularity or desirability estimation and role-playing behavior. The unanticipated order effects in the desirability estimation group place a severe restriction on evaluation of the main hypothesis, yet suggest other post hoc lines of reasoning which are perhaps of greater interest.

The findings with respect to the communality variable appear to be the most straightforward. The *lack* of order effects with this variable suggest that it is not, in itself, an influential contributor to the "set" or strategy which *S*s adopt in role-playing tasks. The meager and inconsistent relationships between indices of communality estimation accuracy and faking success are not supportive of the original hypothesis that this type of estimation is an important parameter of individual differences in role-playing situations. Although not directly compared with the desirability estimation procedure, the communality variable seems clearly less potent as an explanatory construct. A similar lack of explanatory power for the communality variable (as an alternative to social desirability) has been reported in a quite different context (Jackson and Messick, 1962).

The order effects between the social desirability estimation groups suggest that the actual experience of such estimation modifies the subsequent "set" or strategy the *S* adopts in the role-playing task. The fact that this order effect manifests itself primarily in one factorial grouping of role-playing dissimulation scales (*Sd*, *Mp*, *L*) suggests that *S*s who adopt this as part of their approach to role-playing are, in principle, *detectable* by this group of scales. Once the *S* adopts this approach to role-playing, the accuracy with which he is capable of estimating desirability becomes a determinant of the degree of faking success he attains as measured by another factorially distinct group of scales (*Pt* and *A*).

A considerable body of evidence exists (Edwards et al., 1962; Wiggins, 1964; Liberty et al., 1964) which indicates that Edwards' social desirability scale (*SD*) belongs to this last mentioned factorial grouping (*Pt*, *A*) and has similar correlates. It may reasonably be supposed then that individual differences in accuracy will be related to differences in scale score on Edwards' *SD* scale for *S*s who adopt this role-playing strategy. Should subsequent experimentation substantiate this inference it might shed light on some interpretative differences between social desirability measures belonging to the first factorial grouping of the MMPI (*SD*) and those belonging to the third (*Sd*, *Mp*). Subjects who are high on *Sd* but low on *SD* might be *S*s who have adopted, at least in part, the role-playing strategy which emphasizes desirability estimation, but who lack the accuracy of judgement to implement it. Related hypotheses readily occur with respect to other combinations of these two scales.

REFERENCES

Canter, F. M. Simulation on the California Psychological Inventory and the Adjustment of the Simulator. *Journal of Consulting Psychology*, 1963, 27, 253–256.

Cofer, C. N., Chance, June, and Judson, A. J. A Study of Malingering on the MMPI. *Journal of Psychology*, 1949, 27, 491–499.

Crowne, D. P. and Marlowe, D. A New Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349–354.

Edwards, A. L. The Relationship between the Judged Desirability of a Trait and the Probability that the Trait will be Endorsed. *Journal of Applied Psychology*, 1953, 37, 90–93.

Edwards, A. L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden, 1957.

Edwards, A. L., Diers, Carol J., and Walker, J. N. Response Sets and Factor Loadings on Sixty-One Personality Scales. *Journal of Applied Psychology*, 1962, 46, 220–225.

Grayson, N. M. and Olinger, L. B. Simulation of "Normalcy" by Psychiatric Patients on the MMPI. *Journal of Consulting Psychology*, 1957, 21, 73–77.

Hanley, C. Deriving a Measure of Test-Taking Defensiveness. *Journal of Consulting Psychology*, 1957, 21, 391–397.

Hanley, C. Social Desirability and Response Bias in the MMPI. *Journal of Consulting Psychology*, 1961, 25, 13–20.

Hathaway, S. R. and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory Manual* (Revised). New York: The Psychological Corporation, 1951.

Jackson, D. N. and Messick, S. Response Styles on the MMPI: Comparison of Clinical and Normal Samples. *Journal of Abnormal and Social Psychology*, 1962, 65, 285–299.

Liberty, P. G., Lunneborg, C. E., and Atkinson, G. C. Perceptual Defense, Dissimulation, and Response Styles. *Journal of Consulting Psychology*, 1964, 28, 529–537.

McKinley, J. C. and Hathaway, S. R. A Multiphasic Personality Schedule (Minnesota): IV. Psychasthenia. *Journal of Applied Psychology*, 1942, 26, 614–624.

Meehl, P. E. and Hathaway, S. R. The *K* Factor as a Suppressor Variable in the MMPI. *Journal of Applied Psychology*, 1946, 30, 525–564.

Messick, S. and Jackson, D. N. Acquiescence and the Factorial Interpretation of the MMPI. *Psychological Bulletin*, 1961, 58, 299–304.

Rosen, E. Self-Appraisal and Perceived Desirability of MMPI Personality Traits. *Journal of Counseling Psychology*, 1956, 3, 44–51. (a)

Rosen, E. Self-Appraisal, Personal Desirability, and Perceived Social Desirability of Personality Traits. *Journal of Abnormal and Social Psychology*, 1956, 52, 151–158. (b)

Voas, R. B. A Procedure for Reducing the Effects of Slanting Questionnaire Responses toward Social Acceptability. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1958, 18, 337–345.

Walker, J. N. An Examination of the Role of the Experimentally Determined Response Set in Evaluating Edwards' Social Desirability Scale. Ph.D. Dissertation, University of Washington, Seattle, 1961.

Walker, J. N. An Examination of the Role of the Experimentally Determined Response Set in Evaluating Edwards' Social Desirability Scale. *Journal of Consulting Psychology*, 1962, 26, 162–166.

Wiener, D. N. Subtle and Obvious Keys for the MMPI. *Journal of Consulting Psychology*, 1948, 12, 164–170.

Welsh, G. S. Factor Dimensions *A* and *R*. In G. S. Welsh and W. G. Dahlstrom (Eds.) *Basic Readings on the MMPI in Psychology*

and Medicine. Minneapolis: University of Minnesota Press, 1956.

Wiggins, J. S. Interrelationships among MMPI Measures of Dissimulation under Standard and Social Desirability Instructions. *Journal of Consulting Psychology*, 1959, 23, 419–427.

Wiggins, J. S. Strategic, Method and Stylistic Variance in the MMPI. *Psychological Bulletin*, 1962, 59, 224–242.

Wiggins, J. S. Convergences among Stylistic Response Measures from Objective Personality Tests. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 551–562.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

DIMENSIONS OF GROUP INTERACTION: THE COOPERATIVE ANALYSIS OF IDIOSYNCRATIC DESCRIPTIONS OF TRAINING GROUPS[1]

JOHN L. RINN[2]
University of California, Berkeley

THE concern in this study is with the investigation of the perceptual-cognitive parameters underlying the observation, description, and rating of social behaviors. The study also develops a method for measuring the changes over time in descriptions of group behaviors. The method remains close to the data in that the rating scales are cooperatively developed, and in that ratings are performed on "own-statements" rather than on descriptions given by other observers. The study involves certain methodological innovations which have implications for research done within a phenomenological framework. From the viewpoint of graduate education in the social sciences, the procedures promise to contribute to such educational goals as skill in perceiving the verbal and nonverbal behaviors of participants in a group endeavor, skill in formulating perceptions of interpersonal behavior in communicable form, and skill in constructing research instruments for the investigation of perceptual-cognitive data.

Procedures

Members of a seminar in group processes at the University of California, Berkeley, were organized into two training groups of about 15 members each and were encouraged to deal with various developmental tasks which could be solved on the basis of group planning

[1] Supported by funds provided through the Department of Education Research Committee, University of California, Berkeley.

[2] Now in private practice as a psychologist in Berkeley, California.

and group action. The members held secondary school teaching credentials and were preparing themselves for positions as counselors. Their ages ranged from 25 to 35, and each group held approximately the same number of men and women. The educational purposes of this procedure were (1) to provide experiential referents for the theoretical constructs which the students would encounter in reading the literature on group processes, and (2) to provide an opportunity for developing skill in interpersonal observation and interpersonal influence in a group setting.

At the end of each weekly 1½-hour meeting, members were asked to record their "most significant" observations by writing four descriptive statements which completed the following sentence stems: (1) The group . . . , (2) The leader . . . , (3) One member . . . , and (4) I . . . . These statements were collected, collated, and fed back to the group at later meetings so that the members could compare member interests and so that they could become more sensitive to the variations among members in perceptual focusing and communication style. These feedback materials frequently led to a discussion of topics such as the extent to which different statements were descriptions of the same event, or how difficult it was to clarify the relationship between what a person intends to say, what he says, and what others think he says.

Near the end of the semester, the instructor introduced the suggestion that these descriptive statements could be used to assess the directions of group development, and that the members could approach this question as a cooperative research task. One of the groups rejected this suggestion, and the other accepted it and went to work on it. The balance of this paper, then, refers to the data provided by one 15-member training group.

Upon acceptance of the research task, one meeting was devoted to the consideration of categories into which statements could be fitted and, particularly, of the dimensions along which it was hypothesized that changes over time might have taken place. An important criterion was that all members felt that classification of statements in terms of each category was psychologically feasible.

The result of this step was a set of seven scales which could be used to rate the statements which had been collected over a period of eleven meetings. Each of the scales turned out to be bipolar in nature, and a simple seven-step rating sheet was devised which of-

fered a neutral point if the scale was judged to have no relevance for a particular statement. The scales were defined in bipolar terms as follows:

1. personal-impersonal
2. internal-external
3. participation-nonparticipation
4. interest-apathy
5. sensitivity-insensitivity
6. dominance-submission
7. satisfaction-dissatisfaction.

Prior to the last meeting of the group, each member was given his own set of statements to rate, the statements having been separated and coded so that the actual ordering in time was not indicated. The purpose of this step, of course, was to minimize the effect of any preconceptions on the part of members as to which scales should show change over time. The effect of memory could not be eliminated, but it does not seem reasonable that it could have any systematic effect when it is noted that $7 \times 4 \times 11$ separate ratings were involved for each person.

Analysis of Data

Scale Time Trends

To determine whether or not discernable time trends existed in the data, the mean ratings for each session for each type of statement were calculated. It should be noted that the averages were taken for a single scale (e.g., personal-impersonal) and for the same meeting but that the statements which were rated were not identical, each person rating his own statements. Thus, although each statement was a response to a common stimulus stem (e.g., "The group . . ."), the selection of content was idiosyncratic. The appearance of time trends, therefore, would be rather strong evidence that group phenomena were potent enough to overcome both the interperson variance due to perceptual selectivity and that due to differences in rating-scale response.

The 28 mean ratings (7×4) were first examined "by eye" in their graphical form, and smoothed averages were also drawn. Visually, it appeared that several trends were in evidence but since

the scales showed considerable covariance, factor analytic procedures were called upon.

Correlational and Factor Analysis

Each of the four statement categories was treated as a separate social stimulus, and in each case a three-dimensional array of data was available consisting of 15 subjects by seven scales by 11 occasions. This array was first collapsed into a two-dimensional matrix by taking the mean rating over subjects of each scale on each occasion. Intercorrelations of the seven scales were thus based on mean ratings of 15 subjects over 11 occasions. The appropriateness of the procedure depended on the assumption that there were no consistent biases among the subjects concerning how the scales were to be applied to the stimulus statements, and this assumption seemed to be congruent with the fact that the scales were cooperatively developed by the subject-raters. The correlation matrices are presented in Table 1.

The presence of different patterns among the four correlation matrices suggested the possibility that scale interrelationships were to some extent a function of the social object that was being described. Therefore, the four matrices were subjected to separate centroid factor analyses, and an attempt was made to interpret the factors.

The unrotated centroid factor loadings are presented in Table 2. The four factor matrices were rotated orthogonally by hand in terms of the two criteria of simple structure and comparability among the four solutions. Table 3 gives the rotated loadings of all factors. Factors were not rotated unless they contained loadings of .50 or more. Thus, only the first two centroid factors were rotated except for the "One member" solution. This solution was the least satisfactory, although it appears to reflect reasonably well the correlation matrix from which it was derived. It retained some sizable loadings on the third factor which were judged to be specific to the particular stimulus object involved.

The visual record of this operation can be seen in Figure 1 where the graphs of Factor I against Factor II are given for the four stimulus objects. It can be seen that the rotational criteria were reasonably well met, and the similarities in the four solutions were

TABLE 1

Correlation Matrices: Seven Behavior Rating Scales

	"The group"						"The leader"					
Scale	1	2	3	4	5	6	1	2	3	4	5	6
2	55						82					
3	−21	13					82	74				
4	−05	24	58				43	09	24			
5	−16	16	76	73			−25	−49	−50	38		
6	−14	−13	61	38	42		−24	−13	−37	15	58	
7	−18	−03	71	83	80	49	05	−06	09	50	53	44

	"One member"						"I"					
Scale	1	2	3	4	5	6	1	2	3	4	5	6
2	39						39					
3	−13	−30					17	36				
4	−11	22	42				20	40	62			
5	18	−02	−14	−02			60	61	41	33		
6	17	19	17	63	−44		36	13	54	37	36	
7	−63	−18	−14	18	09	−03	−04	04	48	75	36	50

Note—Decimals omitted

TABLE 2

Unrotated Centroid Factor Loadings

	"The group"			"The leader"		
Scale	I	II	III	I	II[a]	III
1. personal-impersonal	09	73	05	−63	70	−19
2. internal-external	35	68	12	−73	51	26
3. participation-nonparticipation	79	−33	−24	−72	53	−25
4. interest-apathy	83	−13	29	22	65	−39
5. sensitivity-insensitivity	83	−29	19	81	24	−09
6. dominance-submission	53	−34	−39	60	29	38
7. satisfaction-dissatisfaction	82	−44	28	50	59	−14
$\sum a^2$	309	151	43	276	194	49

	"One member"			"I"		
Scale	I	II	III	I	II	III
1. personal-impersonal	23	67	43	51	50	24
2. internal-external	32	45	18	56	45	−36
3. participation-nonparticipation	14	−30	−29	71	−27	−17
4. interest-apathy	89	−37	−27	76	−34	−27
5. sensitivity-insensitivity	04	−24	63	73	43	21
6. dominance-submission	60	28	−62	62	−21	33
7. satisfaction-dissatisfaction	−03	−56	−15	63	−56	17
$\sum a^2$	133	133	118	297	118	47

Note—Decimals omitted
[a]Extracted first

TABLE 3

Rotated Factor Loadings

Scale	"The group"			"The leader"		
	I	II	III	I	II	III
1. personal-impersonal	−19	71	05	05	90	−19
2. internal-external	07	76	12	−16	90	26
3. participation-nonparticipation	85	−01	−24	−15	85	−25
4. interest-apathy	82	19	29	62	23	−39
5. sensitivity-insensitivity	88	04	19	73	−39	−09
6. dominance-submission	62	−12	−39	63	−16	38
7. satisfaction-dissatisfaction	93	−10	28	77	03	−14
$\sum a^2$	346	114	43	196	176	49

Scale	"One member"			"I"		
	I	II	III	I	II	III
1. personal-impersonal	−02	79	−27	06	71	24
2. internal-external	18	53	−14	13	71	−36
3. participation-nonparticipation	26	−30	17	70	27	−17
4. interest-apathy	80	02	57	79	24	−27
5. sensitivity-insensitivity	−40	30	45	28	80	21
6. dominance-submission	88	04	−20	60	25	33
7. satisfaction-dissatisfaction	00	−44	38	84	00	17
$\sum a^2$	167	128	83	228	184	47

Note—Decimals omitted

such that an interpretation was attempted. The analysis below can be followed by referring to the graphical or tabular solutions.

Interpretation of Factors

The most striking feature of the four factor plots was the appearance of two comparable factors from each set of scales. Thus, one factor was defined by Scales 1 and 2 on all four plots and can be referred to as the *degree of personal reference* factor. That is, a social object may be described at the personal, internal level, or it may be described in terms of those characteristics which are external and can be objectified in an impersonal manner.

The second factor was identified as *degree of involvement* and tended to be defined by the other five scales. For example, "The group" was seen either as being composed of interested, assertive, satisfied participants or of members who tended to withdraw in apathetic dissatisfaction.

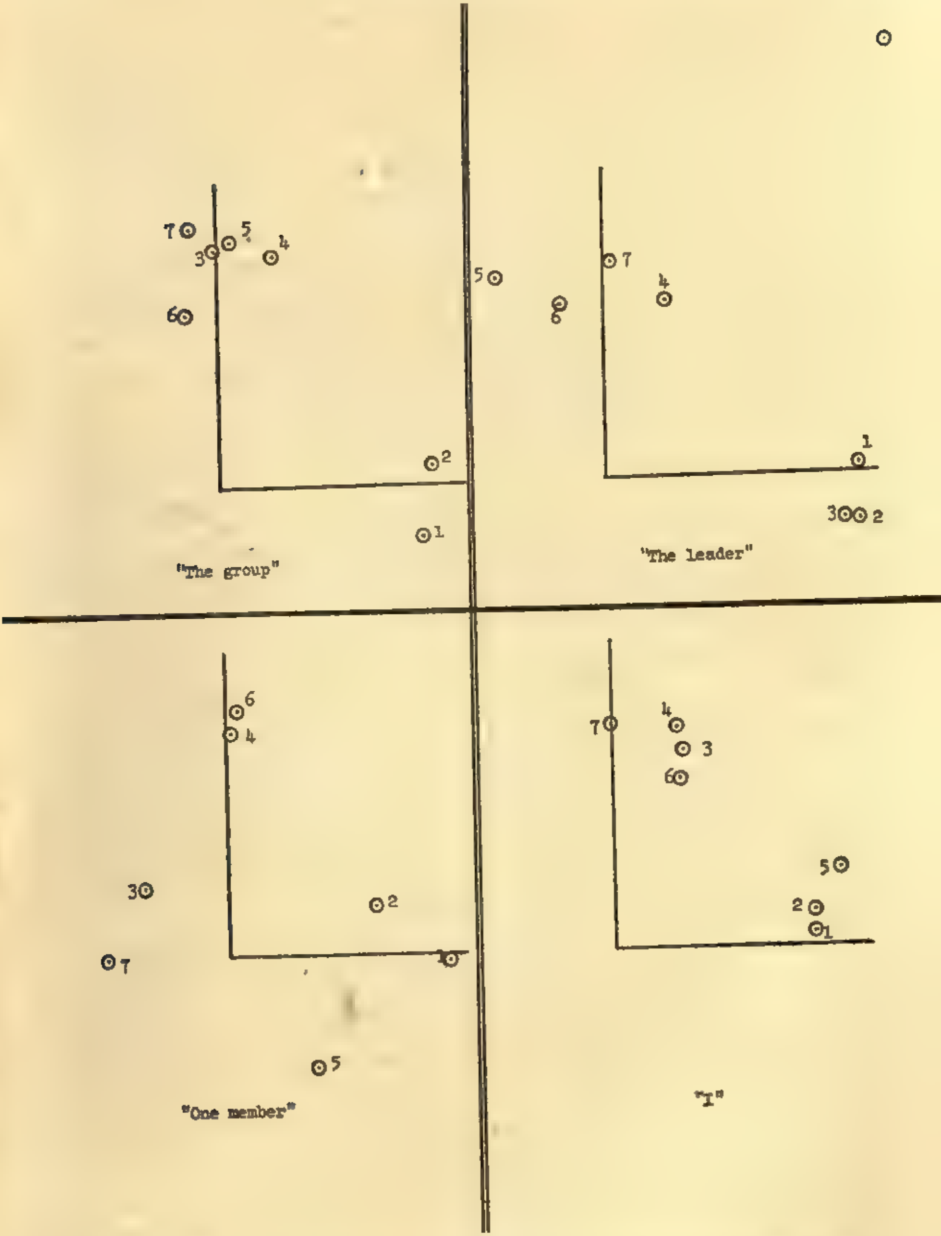


Figure 1. Plots of First Two Rotated Factors for Various Stimulus Objects.

Another important feature of the four factor solutions was the manner in which some scales switched their factor association depending on which social object was being rated. Thus, for the rat-

ings of "The leader" statements, the scales covaried as described above with the important exception that the participation-nonparticipation scale (Scale 3) was now associated with the *personal reference* factor. The implication seems to be that when the group leader participated, it was on a personal level—he expressed personal opinions, or gave subjective impressions of events, or referred to internal events in himself or the group. For the leader, then, participation in the group was an interpersonal process. Discussion of impersonal (intellectual) or external (out-of-field) topics by him was seen as nonparticipative in the interpersonal sense.

For the "I"-statement ratings, the two-factor pattern again had one deviant element. This time, it was the sensitivity-insensitivity scale (Scale 5) which became associated with the *personal reference* factor. This says that more sensitivity was involved in referring to one's internal self than to one's external self. Although this is not a surprising relationship, it points up the fact that members did not associate sensitivity with personal expressions of the leader or other members.

The factor pattern derived from "One member" statements is least similar to the other three patterns. One explanation would be that the variability of the social object (*any* member in each group meeting) led to a more complex set of associations than was evoked by the other descriptions. However, this explanation would not account for the lack of complexity among the parameters of the "I" statements since these statements also included references to all group members. Apparently there is greater commonality in self-perceptions than there is in other-perceptions. At any rate, the *personal reference* factor for these statements carried a negative loading for Scale 7 which meant that members were seen as dissatisfied when they expressed personal content rather than satisfied; or alternatively, statements were more often written about the personal references of other members when those references were negative than when they were positive.

A negative loading also appeared on the *involvement* factor, this time on the senstivity-insensitivity scale. Thus the involvement of other members was perceived as varying from insensitive assertiveness to self-conscious restraint and contrasted with the more sensitive interactions of the leader and the group as a whole. This pattern is suggestive of the distinction which has been made by Carter

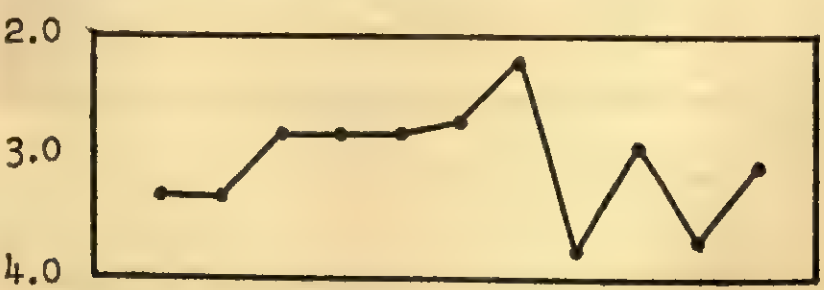
(1954) and others between "self-oriented" and "group-oriented" behaviors.

Factorial Time Trends

In the previous section on scale trends over time, it was noted that the scales had been examined graphically and that because of the appearance of considerable covariance among the seven scales, it had been decided to factor analyze the scales before attempting to assess meeting-to-meeting trends. Thus the factor analyses revealed that a major portion of the scale variance over time could be accounted for in terms of two factors, one a *content* factor and the other a *process* factor. This finding permitted the analysis of time trends to proceed in terms of factorial constructs characterize by greater parsimony and reliability than was held by the separate scales.

To assess the time trends of the *personal reference* factor, new average ratings of each social object were calculated for each meeting based on the summed ratings of those scales which had high loadings on that factor. This meant that the average ratings of statements on "The group" and on "One member" were based on Scales 1 and 2; "The leader," on Scales 1, 2, and 3; and "I," on Scales 1, 2, and 5. Graphs of these averages are presented in Figure 2.

The same procedure was applied to the *involvement* factor, and in this case the factorial rating of "The group" was derived from Scales 3, 4, 5, 6, and 7; "The leader," from Scales 4, 5, 6, and 7; "One member," from Scales 4 and 6; and "I," from Scales 3, 4, 6, and 7. Graphs of these trends are presented in Figure 3.



"The group"

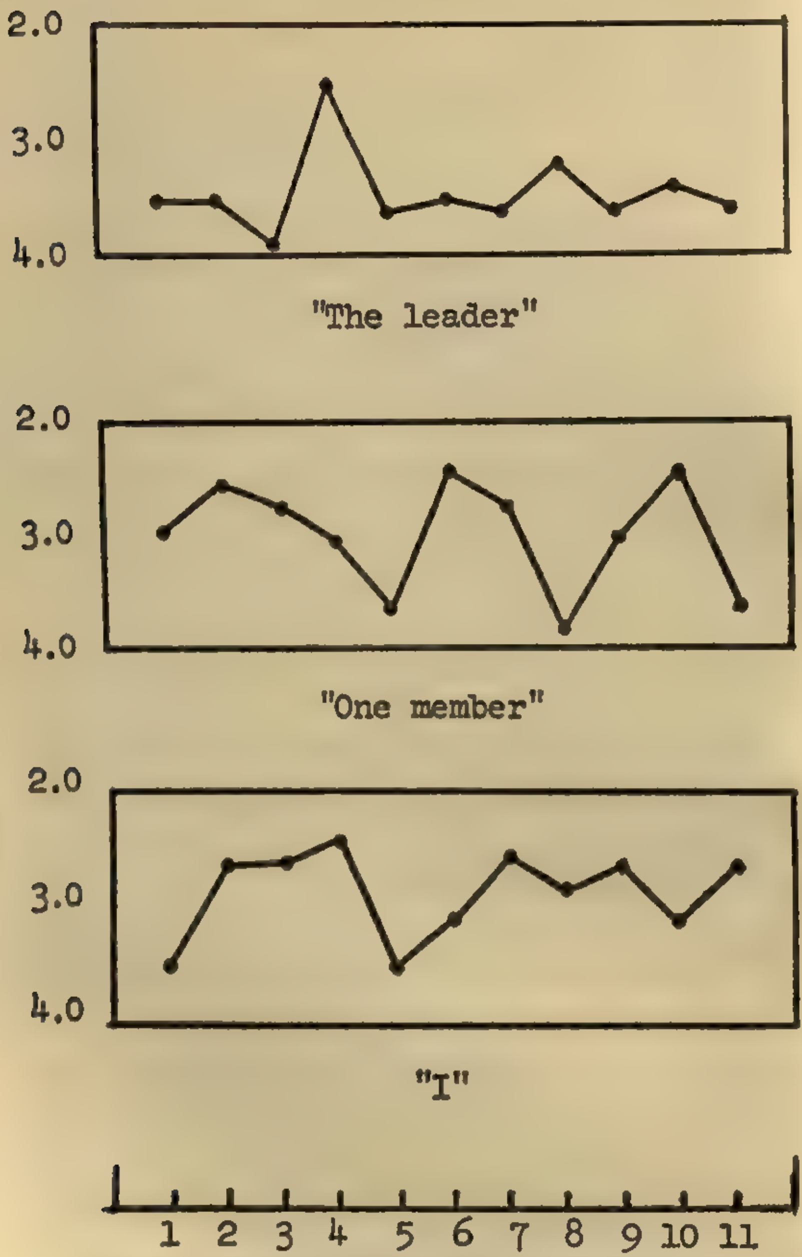
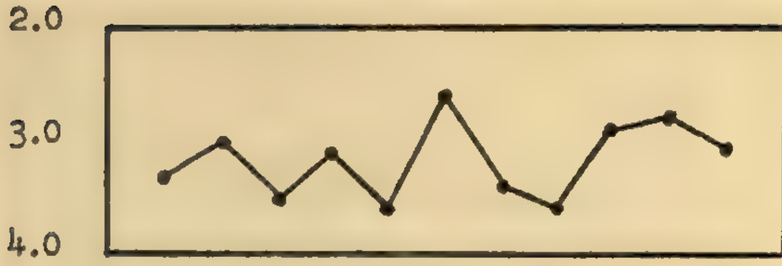


Figure 2. Factorial Time Trends: Personal Reference Factor.



"The group"

"The leader"

"One member"

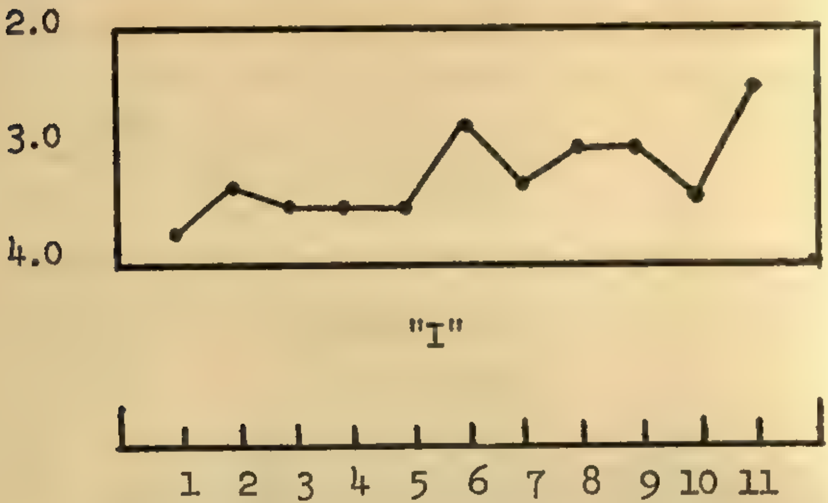


Figure 3. Factorial Time Trends: Involvement Factor.

No adequate statistic is available to describe data of this nature. However, a subjective analysis was made that led to descriptions of group movement which can be compared with results of other studies. This analysis resulted in the following descriptive summary of this training group:

(1) While involvement by the group as a whole was characterized by an approach-withdrawal cycle throughout the life of the group, personal references increased gradually towards a peak at the seventh session and then oscillated for the remaining sessions.

(2) The leader's behavior was constant throughout except for a peak of personal referencing during the fourth session.

(3) While other members' involvement showed little meeting-to-meeting change, their personal references oscillated strongly during the second half of the life of the group.

(4) Persons saw themselves as gradually becoming more and more involved during the life of the group. After the first session, their personal references were at a high level throughout, except during the fifth and sixth sessions.

Discussion

Methodology

An important formal dimension of social-phenomena observer

systems has to do with the nature of the units which are used to identify and describe temporal aspects of behavior. Heyns and Lippitt (1954) have pointed to the difference between natural and imposed units and have noted that observational techniques were moving away from a concern with exhaustiveness and high inter-rater reliability and toward selectiveness, inferential observation, and theoretical relevance. More recently, Barker (1963) has distinguished between subject-generated *behavior units* whose boundaries are independent of the operations of the investigator and *behavior tesserae* which are imposed upon the behavior stream in accordance with an investigator's scientific aims and preconceptions.

In the present study, the basic behavior descriptions were made by members themselves without the imposition of a predetermined category system. Although statements were generated from four sentence stems as a means of focusing attention upon each relevant social object in the environment, members were free to make comments in terms of their own criteria for unit formation. These comments, then, are examples of *natural units of content,* although the attention of the observers was directed rather than free.

A related characteristic of the present data system has to do with the criteria of selectivity which participants are presumed to have employed. In a broad sense the principle of selection may be said to be that of *perceptual significance.* As Hare (1962, p. 81) has pointed out, social perception is a complex phenomenon which is a function of cultural attitudes, the perceiver's personality, the personality of the perceived object, unique situational factors, and the action at the moment. Efforts to isolate the effects of these various factors have not met with great success, and from a holistic point of view (e.g., Lecky, 1951; Combs and Snygg, 1959), such isolation is of dubious worth even when accomplished. The experimenter's values and purposes are crucial at this point. He must ask if it is more important to test how well observers can use the experimenter's category system or to permit observers to respond in terms of their own frame of reference. Should he guide observers to a predetermined conceptual viewpoint or permit them to discover their own perceptual commonality? Obviously, different answers to these questions should lead to different research designs, and examples of both should be found in the literature. However, the dependence of

social scientists upon the experimental methods of the physical scientists has meant an absorption of their values also, and most research in the behavioral sciences has tried to justify itself by the standard of objectivity. Fortunately, the recent development of new methodologies promises to free social science research from the tyranny of the correlation coefficient (Coombs, 1963; Shepard, 1962), and nonmetric multidimensional analyses of interpersonal data have begun to appear (Rinn, 1963).

Phenomenology and Education

The decision to encourage idiosyncratic responses has both general theoretical grounds and specific educational purposes. Within a phenomenological framework, the explication of phenomena is essentially an inductive process whereby the reality of any phenomenon is approached through a systematic attempt to observe and describe the world of experiential reality as it is presented to the person. This theoretical framework is congruent with the process-oriented approach to group learning which has most recently been summarized by Bradford et al. (1964). The major principles of this approach are a focus on the phenomenal field of the here-and-now and the development of member sensitivities to social phenomena; the discovery and utilization of various methods of inquiry, including involvement, observation, feedback, and experimentation; the development of member skill in the processes of group membership, decision-making, self-assessment, and problem solving; and the transition from a group technology based on suspicion, facade, manipulation, and dependency to one based on trust, openness, permissiveness, and interdependence.

Cooperative Research

It is important to note that the participant-observers in the present small-group, cooperative-learning, process-oriented setting were able to write descriptive statements beginning with four sentence stems and that this much structuring of descriptions was not resented by members who had opposed almost all other attempts to initiate structure into the group's procedures. The advantages of the sentence-stem method are that it elicits responses to each element in the social field, the responses are in a form which is easily adaptable to content analysis procedures, the method interferes mini-

mally with the ongoing process itself, and the method is congruent with educational goals which stress the active development of perceptual, descriptive, and interpersonal skills.

Since the educational process included the value of cooperative development of research procedures, it would have been inconsistent to perform a content analysis of the data using externally derived scales. Thus, the criterion of *participant significance* was applied to the statement-rating process as well as to the observational process which produced the statements. Obviously, other rating scales could still be applied to the descriptive items, but if they are applied by nonparticipating raters, then any interpretation of ratings must be made in nonparticipative terms. Such interpretations can be useful for some purposes, but their relationship to the findings of the present study is that of a distant cousin, twice removed.

A final aspect of the cooperative-research model is the feedback of data to the members of the project. In the present study this feedback took two forms: immediate summaries of member statements, and a final summary of member ratings of statements over the series of eleven meetings. The final summary was mailed to members after the semester ended and consisted of the graphs of class averages of the seven scales for each sentence-stem category (28 graphs). Each member also received his own ratings which he could analyze for time trends or compare with the class averages.

It appears that this final step is an important one if members are truly to learn that the research process can be one of participants operating with each other rather than of being operated upon. Students are not only unlikely to encounter nonmanipulative designs in the research literature, but the idea is likely to remain an idealistic fiction unless they fully participate in the procedures implied by a cooperative philosophy.

Group Dimensions

A major part of the variance associated with ratings of descriptive statements of the various components of the social field can be understood in terms of two orthogonal dimensions: *degree of involvement* (both verbal and emotional) and *degree of personal reference.* It was as if members were asking two sets of questions about the group process. One set included such concerns as: Are we going to work closely together, or keep our distance? Will we be

sensitive enough to give and take, or will the powerful ones take over? Will the interaction be an interesting and satisfying one, or dull and unproductive? These concerns appear similar to those which have been subsumed by Schutz (1958) under the heading *inclusion.* They are also similar to those included in an *involvement* factor which was extracted by Rinn (1963) from free-response descriptions of the early meetings of a graduate course in counseling psychology. As in the present study, categories for a content analysis were supplied by the students themselves. Although categories were not developed cooperatively, the set of statements was divided by each student into several idiosyncratic categories of "similarity." Frequency of joint inclusion in a category for each pair of statements provided a matrix of percentages which was analyzed by Coombs' (1963) nonmetric multidimensional unfolding procedures.

The *personal reference* dimension can be understood as representing a group concern which may be characterized by another set of questions: Are we going to talk about our selves or about our ideas? Will we focus on the here-and-now or the there-and-then? Will we risk personal exposure or keep ourselves out of the spotlight? These concerns have to do with the *content* of interaction rather than the *form* (Hare, 1962, p.12). They suggest that a basic polarity which can split a group is that which divides members into those who accept the self as the central variable in social interaction from those who prefer to explain behavior in terms of environmental or other influences. Benne (1964) has provided an excellent discussion of the manner in which groups struggle to resolve this and other paradoxes of group life.

Time Trends

Involvement. The time-trend analysis of the *involvement* factor in the present study revealed that members saw different involvement patterns for the various social objects. Thus, the temporal stability of the leader contrasted with a gradually increasing involvement of the self and also with a fluctuating cycle of approach and withdrawal by the group as a whole. The stable role of the leader was to some extent a result of deliberate self control, but the increasing self involvement of members would seem to reflect a steady growth in member confidence and spontaneity.

The approach-withdrawal cycle of the group as a whole deserves

special consideration as evidence of a phenomenon which has not previously been observed at the group level. The phenomenon has been utilized at the individual level by the Gestalt therapists Perls, Hefferline, and Goodman (1951), however. Perls sees the rhythm of contact and withdrawal as being a necessary process in the creative adjustment of the self to the world, and he has devised a series of therapeutic techniques whose purpose it is to bring the patient into harmonious adjustment with this rhythm.

If the existence of the ebb-and-flow phenomena should be verified, it would seem to have significance for training-group theory and practice. From a group member's point of view, the phenomenon would appear as a heteronomous event, to use Angyal's (1941) term, which contrasts with autonomous happenings having their source in the person himself. Growth in member competencies is currently conceptualized within a framework which treats heteronomous difficulties as temporary environmental conditions which, as Heider (1958) points out, are commonly diagnosed as due to opportunity or luck. These conditions tend to be seen by group trainers as disturbing the steady growth of member potentials and are dealt with as process problems or blocks to group productivity. An alternative conceptualization implied by the present findings is that the phenomena of intrapersonal growth and interpersonal interaction have different natural patterns. This implies that interpersonal contact and withdrawal should be dealt with as necessary phases of a natural polarity rather than in terms of a valued capacity for involvement and a devalued tendency for resistance and defensiveness. Involvement and detachment would then both be seen to have their unique values, and group trainers could deal with the phenomena descriptively rather than judgmentally.

Personal reference. Turning now to the second factorial time trend, it can be noticed that the personal reference factor showed a different pattern for perceptions of the leader and self perceptions than it did for perceptions of the group and other members. Except for a high point during the fourth session, the leader's contributions stayed at a constant moderate level of personalness. After the first session, members also rated their own contributions at a constant—although slightly higher—level except for the fifth and sixth sessions. The generally higher level of personal reference by "oneself" is easily accounted for in terms of the viewpoint of the observer—

whatever I say is more personal to me than what you say. The dip in the fifth and sixth sessions, however, may be related to the leader's peak in the session a week previous. It may be hypothesized that intense personal referencing by the leader leads to a temporary retreat to impersonal content by members, although this hypothesis should be put to experimental test. If the relationship should be found to hold, its explanation could lie in the fact that the observer of an intense personal expression frequently feels more residual embarrassment and self consciousness than the immediate participant, possibly because the observer tends to be blocked off from active engagements which might serve to resolve his tensions. Goffman (1959) suggests that the observer's discomfort may be due to a disturbance of expectations regarding the actor's role, and "that even sympathetic audiences can be momentarily disturbed, shocked, and weakened in their faith by the discovery of a picayune discrepancy in the impressions presented to them" (Goffman, 1959, p. 51). By this interpretation, members in the present study expected the leader to maintain a low level of personal reference, were discomfitted when he did not, and turned for a time to an impersonal content level.

The dips on the *personal reference* factor for ratings of other members can mostly be accounted for in terms of the above explanation. Thus, the dip which occurred during the fifth session followed the leader's peak in the fourth meeting, while another dip in the eighth session followed a peak by the group in the seventh meeting.

The *degree of personal reference* for the group as a whole followed the interesting pattern of gradually increasing for seven sessions and then fluctuating back and forth for the final four meetings. Having accepted the goals of involvement and personal reference, it appears that this group reached its maximal level of personal intensity during the seventh meeting and then began a "clinging-off" period in preparation for a return to less process-oriented work environments.

Conclusion

If one is interested in determining the frequency with which statements of various types are made, or in getting an "objective" record of member behaviors in a group, then the methodology which is

illustrated in the present study is inappropriate. However, if one takes a phenomenological approach to the study of social behavior, or if one is concerned with the felt importance of events to persons rather than with their frequency, then it is sensible to permit each subject to select those events that have significance for him. The record that is obtained under the latter condition is one of subjective value rather than one of objective frequency. When such data are summed and averaged, as in the present study, any relationships that can be shown to exist have the promise of truly representing group phenomena. Although the hypotheses that are derived from such relationships are likely to be weak, they are also likely to have great generality.

REFERENCES

Angyal, A. *Foundations for a Science of Personality.* New York: Harvard University Press, 1941.

Barker, R. The Stream of Behavior as an Empirical Problem. In R. Barker (Ed.), *The Stream of Behavior.* New York: Appleton-Century-Crofts, 1963. Pp. 1–22.

Benne, K. From Polarization to Paradox. In Bradford et al., *T-group Theory and Laboratory Method.* New York: John Wiley & Sons, 1964. Pp. 216–247.

Bradford, L., Gibb, J., and Benne, K. *T-group Theory and Laboratory Method.* New York: John Wiley & Sons, 1964.

Carter, L. F. Evaluating the Performance of Individuals as Members of Small groups. *Personnel Psychology,* 1954, 7, 477–484.

Combs, A. and Snygg, D. *Individual Behavior: A Perceptual Approach to Behavior.* New York: Harper, 1959.

Coombs, C. *A Theory of Data.* New York: John Wiley & Sons, 1963.

Goffman, E. *The Presentation of Self in Everyday Life.* New York: Doubleday & Co., 1959.

Hare, A. P. *Handbook of Small Group Research.* New York: Free Press of Glencoe, 1962.

Heider, F. *The Psychology of Interpersonal Relations.* New York: John Wiley & Sons, 1958.

Heyns, R. and Lippitt, R. Systematic Observational Techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology.* Cambridge: Addison-Wesley, 1954, Vol. 1. Pp. 370–404.

Lecky, P. *Self Consistency: A Theory of Personality.* New York: Island Press, 1951.

Perls, F., Hefferline, R., and Goodman, P. *Gestalt Therapy.* New York: Julian Press, 1951.

Rinn, J. Group Behavior Descriptions: A Nonmetric Multidimensional Analysis. *Journal of Abnormal and Social Psychology,* 1963, 67, 173–176.

Schutz, W. *FIRO: A Three-Dimensional Theory of Interpersonal Behavior.* New York: Holt, 1958.

Shepard, R. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. *Psychometrika,* 1962, 27, 125–140.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

SECULAR TRENDS IN AN ADJECTIVE CHECKLIST

WILLEM K. B. HOFSTEE[1]
Royal Netherlands Navy[2]

When forced-choice personality inventories are constructed on the basis of item indices (such as favorability or popularity) as obtained through administration of the items in some rating-scale format, one assumption implicitly made by most test constructors is that these indices are not biased appreciably by the position in which the items occurred. More generally, such an assumption is implicit in any selection process whereby items are chosen on the basis of certain characteristics, and where position has not been controlled in assessing these characteristics.

Gordon (1952), among others, demonstrated that position does influence response to items in long personality inventories. By systematically varying item position, he was able to show that subjects tend to respond in a more desirable manner as the list proceeds. One of Gordon's conclusions was that some method of rotating the items would be desirable when long lists of items are presented for the purpose of obtaining preference indices for use in the construction of forced-choice inventories.

In a recent study (Stricker, 1964) that reviews most of the literature on secular trends (cf. Loevinger, 1957, p. 676 ff.), a similar finding with respect to criticalness is reported: the extent to which an item measured such a response style was found to be related to its location in the test, while the content validity of the item was not predictable from its location.

[1] The author wishes to thank Drs. Lawrence Stricker and Richard Melton for their comments on this paper.

[2] The present investigation was completed while the author was a fellow of the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) at Educational Testing Service.

The term "secular trends," which was originally (Loevinger, 1957) proposed to account for systematic changes that take place upon readministration of a test or its parallel form, and which was subsequently broadened to subsume such changes within a single administration (Stricker, 1964), could be further expanded to include even shorter-term trends such as order and context effects, described by Cowen and Stiller (1959). The authors administered a list of 30 neutral trait-descriptive adjectives in counterbalanced order; subjects were to rate the traits for desirability. Context effects appeared in that positive adjectives rated first were seen as less socially desirable than in the setting in which item indices had been obtained; neutral adjectives were seen as relatively more desirable when rated first. Order effects appeared to the extent that positive adjectives became more desirable when rated after neutrals, while the reverse order decreased the desirability of neutral items.

Psychometrically, the situations just described are similar in that item response is to some extent dependent upon whether or not subjects have answered other items of the same sort; whether the basis for such sequential redundancy be fatigue or boredom, changes in the subjects' cognitive or emotional approach to the test, or changes in adaption level (Cowen and Stiller, 1959). It might be profitable theoretically to reserve the term "secular trends" for the effects of any such changes that are produced by experience with the test itself, as opposed to learning outside the testing situation.

The present research was undertaken to assess the influence of secular trends upon response to a list of trait-descriptive adjectives, insofar as such trends result in biased estimates of popularity indices. Two hypotheses were tested: (a) popularity indices of items which occurred later in the list would turn out to be overestimated in the case of positive adjectives, and underestimated in the case of negative traits; (b) the popularity value of a trait is influenced by the popularity of the trait immediately preceding it.

Method

A list of 180 trait-descriptive adjectives was administered to 120 Dutch military subjects. Instructions were to indicate on a five-point scale the extent to which a trait was applicable. Responses

were weighted 1 to 5 and averaged for each trait to provide popularity indices (PI).

On the basis of these PI, forced-choice formats were constructed. These formats consisted of tetrads, each having a pair of popular and a pair of less popular traits. The two items forming a pair were closely matched on PI. Instructions for the forced-choice format were to indicate, for each tetrad, the trait that was "most applicable" and the trait that was "least applicable."

Two forms were constructed. Form A consisted of 20 tetrads; it was administered to a sample of enlisted men in the Netherlands Navy, of which 100 were chosen randomly for the analysis. Form B consisted of 36 tetrads and was given to a sample of Dutch army men, of which 476 were selected randomly.

For each item (80 in Form A, 144 in Form B), "most applicable" and "least applicable" responses were tallied, and the latter number was subtracted from the first. The item index thus obtained will be referred to as DI (difference index); it is, however, merely another popularity index, obtained in a different context.

For each pair of items which were matched on PI, a comparison was made between the two DI, and it was attempted to predict the direction of the difference between the two DI on the basis of hypotheses (a) and (b). Predictions were (a) that for a pair of popular items (items with PI > 3.00), the member that occurred later in the original list would have the lower DI (because its PI would be inflated); and that for a pair of unpopular items (PI < 3.00), the member that occurred later in the rating list would have the higher DI; and (b) that for each pair of items, that were matched on PI, the direction of the difference between the two DI would be systematically related to the direction of the difference between the PI of the single items which immediately preceded the members in the original list.

Results and Discussion

Hypothesis (a), concerning a growing set to respond desirably in a long list of items, was not substantiated: of the 40 predictions that were made for Form A, 21 were in the right direction and 19 were wrong; for the 72 pairs of Form B, 38 predictions were right and 34 were wrong. The position effect as demonstrated by Gordon (1952) does not seem to have been particularly influential. There

may be at least two reasons for the lack of practical significance of this secular trend in the present case. In the first place the present list, even though containing numerically as many items as Gordon's, was in all likelihood shorter as far as testing time is concerned: adjectives require less reading time than statements. Second, the influence of other effects may have been so powerful as to obscure position trend.

Hypothesis (b), representing the influence of the previous item's desirability on PI, was clearly confirmed by the results, to the extent that the PI of the members that were preceded by more desirable items turned out to be overestimated; consequently, these members had the lower DI within a matched pair. With this direction in mind, 25 out of 39 predictions were right for Form A ($\chi^2 = 3.10$, $p < .10$), and 49 out of 71 were right for Form B ($\chi^2 = 10.27$, $p < .01$). In both cases, one prediction could not be made because of a tie in PI.

In view of the indirect method employed, it is somewhat surprising that the effect of the preceding item upon trait popularity could be so clearly demonstrated. The effect seems to be of contamination: the PI of an item is biased in the direction of the preceding item. Subjects appear to perseverate in responding to a list of traits, whether such perseveration is of motor or of more central nature.

In a sense, the present findings are opposite to the order effects found by Cowen and Stiller (1959). While in their study a contrast effect appeared, the present results point toward assimilation. Presumably, the difference is one of design: while Cowen and Stiller used lists that were very homogeneous with respect to social desirability, the present study employed a very heterogeneous list. It seems plausible that a homogeneous list should produce contrasting tendencies, while a sequentially heterogeneous list should elicit perseveration.

The results of this investigation illustrate once more the necessity of taking secular trends into account; if the broader definition here proposed is accepted, it would be hard to maintain that "the 'theory of the first test' can ignore them" (Loevinger, 1957; p. 678). Furthermore, the study of secular trend may well contribute to further insight into what subjects are actually doing when they fill out a personality inventory. At least in personality testing,

the notion of an isolated response to an isolated item seems to become less and less realistic.

Summary

Secular trends in tests were redefined as effects of systematic changes in response to test items resulting from experience with similar items, rather than from learning outside the testing situation. In the present study, two such trends were investigated in a list of 180 trait-descriptive adjectives: a trend for items to be responded to more desirably as the list continues, and a trend for items to be responded to more or less desirability depending upon the desirability of the preceding term. The method consisted of comparing item popularities, as found in the original format, with the popularities of the same items presented in forced-choice format: if item popularities had been biased through the operation of secular trends in the original list, such bias should to some extent be detectable in the responses to the forced-choice format. The results showed no evidence for the first trend; with respect to the second, it was clearly demonstrated that items in the original list were responded to more desirably as the preceding item was more desirable. The importance of controlling and studying secular trends was discussed.

REFERENCES

Cowen, E. L. and Stiller, A. The Social Desirability of Trait Descriptive Items: Order and Context Effects. *Canadian Journal of Psychology*, 1959, 13, 193–199.

Gordon, L. V. The Effect of Position on the Preference Value of Personality Items. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1952, 12, 669–676.

Loevinger, Jane. Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, 1957, 3, 635–694 (Monograph Supplement, 1957, No. 9–V3).

Stricker, L. J. Difficulty and Other Correlates of Criticalness Response Style at the Item Level. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1965, 25, 683–706.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

INTEGRATIVE COMPLEXITY: ITS MEASUREMENT AND RELATION TO CREATIVITY[1]

BRUCE W. TUCKMAN
Rutgers; The State University, N. J.

Harvey, Hunt, and Schroder's (1961) classification of individuals according to the level of integrative complexity of their personality structure has led to the accurate prediction of the individual's diversity of perceptions (Driver, 1962), reaction to attitude refutation (Streufert, 1962), and attitude change in a sensory deprived and isolated environment (Suedfeld, 1964). The classification system, when used to compare homogeneous groups, has made possible the prediction of sensitivity to feedback (Lawrence, 1962) and group structure and performance on a variety of measures in a simulated environment (Tuckman, 1964). This evidence provides validity for the classification scheme and its underlying theory, as well as the measure used for identifying the individual's level of integrative complexity. The measure used for this purpose has been the Sentence Completion Test (SC) developed by Schroder and Streufert (1962), a projective instrument, which can be considered to have been validated by the above studies.

The purpose of this study was to evaluate an objectively-scored device to measure integrative complexity by determining the extent to which it was useful in predicting creativity, a characteristic

---

[1] From Bureau of Medicine and Surgery, Navy Department, Research Task MR005.12-2005.01, Subtask 1. The opinions and statements contained herein are the private ones of the writer and are not to be construed as official or reflecting the view of the Navy Department or the Naval Service at large.

The author would like to thank Mr. Robert Nye for his help in collecting the data and Dr. Irwin Altman for his valuable suggestions in analyzing and interpreting the data.

implied in the concept of integrative complexity, and to determine further whether this new instrument classified individuals similarly to the SC. Since the SC, as a projective instrument, is difficult to score reliably, and scoring must be preceded by the training of scorers which is arduous and time-consuming, an objectively-scored measure of integrative complexity would be useful. A further advantage would be gained in dealing with persons of limited verbal ability who respond tersely and incompletely to open-ended questions.

Four nodal systems of integrative complexity have been described in detail by Harvey, *et al.* (1961).

System I. At the lowest level of integrative complexity, the rules or schemata for categorizing stimuli are highly fixed and simple. Ambiguity is not tolerated and simple schemata, norms, or authorities help the individual to structure his environment in a complete and unyielding way. System I individuals are characterized by categorical, black-white thinking, minimization of conflict and avoidance of ambiguity, self definition in terms of external anchors, preservation of standards and minimization of alternatives, and overgeneralization of fixed approaches or stereotypes.

System II. At a level of integrative complexity somewhat above that of System I, the schemata for categorizing stimuli are still relatively simple, but more alternatives are perceived than in the case of System I functioning. The System II individual perceives his world against a background of self vs. other (representing the separation of self and other which is beyond the integrative complexity of System I), and accepts self while rejecting other. This leads to an absolutistic orientation toward others who, when seen in a position of potential control are "warded off;" one experiences conflict when external standards are imposed on the self. This detached, antidependency characterization has been termed a negatively independent orientation.

System III. At this higher level of integrative complexity the schemata for "reading" the environment are more flexible and more alternatives are perceived. Not only is the self highly differentiated but other people are equally differentiated. This latter fact enables the System III to be highly sensitive to others and to attempt to match his perceptions to those of others. That is, he is highly capable of putting himself in the role of others and perceiv-

ing himself as others perceive him. The orientation, then, is toward the maintenance of close interpersonal relationships and rejection is threatening.

System IV. At the highest level of integrative complexity a diverse world filled with many alternatives is perceived. The System IV individual uses highly complex and flexible schemata for reading his environment and those in it. Interpersonally, this individual is highly autonomous and reacts to people as a source of information. The System IV person generates a large variety of alternative interpretations of environmental events and can thus react to the subtleties of his environment with appropriate and novel responses.

Problem

Generalizing from the descriptions of the four systems, it was decided that the following criteria would be used in evaluating an objectively scorable measure of integrative complexity, henceforth designated as the Interpersonal Topical Inventory (ITI):[2]

1). The system classification of an individual based on his profile of scores on the ITI should make possible the prediction of his performance on a battery of tests of creativity following the assumption that the more integratively complex the individual, the more creative he will be.

The anticipated relationship between integrative complexity and creativity was based on the fact that the former is a measure of the extent to which the individual separates (i.e., differentiates) and recombines (i.e., integrates) inputs in a variety of ways, while the latter is a measure of the extent to which the individual's thinking embraces novelty and speculation and leads to the production of unusual but relevant responses. Integrative complexity describes the structure of the individual's life space in terms of its degree of differentiation and integration while creativity describes the characteristic responses produced as a function of the individual's level of integrative complexity. The more alternatives the individual apprehends in his thinking, the more likely he is to produce novel, creative responses. Moreover, the less anchored the individual's

---

[2] This is not a test of validity in the traditional psychometric sense. Rather, it is an attempt to evaluate the research usefulness of a new, objectively-scored instrument for classifying people. Consequently, test-retest reliability and internal analyses characteristic of a thorough-going psychometric effort will not be presented.

thoughts are in the external situation and external criteria, the more likely they are to be creative. Since integratively complex persons produce more alternatives and are less externally constrained, they were expected to be more creative, i.e., produce more creative responses.

If the ITI is a useful measure of integrative complexity, individuals classified by the ITI into systems should be rank ordered System IV, System III, System II, System I on creativity with IV's most creative. The system classification of the individual based on his scores on the SC should lead to the same rank-ordering of the system groups on creative performance as that for the ITI.

2. The system classification of individuals based on ITI profile scores should correspond to classification on the SC scores. If classification using the ITI corresponds to classification using the SC, the more objective ITI measure will then be considered a potentially useful substitute for the SC as a measure of integrative complexity.

Method

Subjects

Subjects were 126 Navy enlisted men from the U. S. Naval Service School Command, Bainbridge and the U. S. Naval Station, Washington, D. C.[3] The median age of *Ss* was 18 with a range of 17 to 24. Seventy-nine per cent of the *Ss* were high school graduates while 8 per cent had, in addition, some college experience. None had a college degree.

Intelligence of the subject population was assessed by the Navy General Classification Test (GCT) and only individuals having GCT scores of 53 (IQ = about 106) and over were included in this study. The 126 *Ss* had a median GCT score of 59 with a range of 53 to 71.

The Predictor Measure: The Interpersonal Topical Inventory

The Interpersonal Topical Inventory of integrative complexity (ITI) is a forced-choice instrument in which *S* is asked to choose

---

[3] The author would like to thank CDR. A. D. Garvin, USN, commanding officer of the U. S. Naval Service School Command, Bainbridge, Md. and Chief Ships Clerk W-3 J. F. Cullinan, separations officer at the U. S. Naval Station, Washington, D. C. for providing the subjects used in this study.

one of a pair of items that best represents his feeling about or reaction to an interpersonal topic. The topics are: (a) when criticized, (b) when in doubt, (c) when a friend acts differently toward you, (d) beliefs about people in general, (e) feelings about leaders, (f) feelings about rules. These topics are meant to confront the individuals with interpersonal conflict, ambiguity, and the imposition of control. For each topic there are six pairs of alternatives; *S* must pick one from each pair. Each member of a pair represents a typical response for a different system. All possible paired combinations of the four systems go to make up the six pairs.

S is given four scores, each representing his total number of choices for each of the four systems. The maximum score for any system is 18. Based on the frequency distribution of system scores for the 126 *S*s used in this study plus another 85 *S*s run previously for other purposes, half-decile ranges (each containing five percentiles) were identified and each of the four scores for each *S* was replaced by a percentile score and a half-decile score. *S*s scoring in any one of the 8 highest half-deciles (i.e., scoring at the 61st. percentile or better) on one of the four systems and lower on the other three were assigned to that system.[4] Thus, system assignment was on a relative rather than an absolute basis. On this basis, 31 *S*s were classified as System I, 26 as System II, 22 as System III, and 30 as System IV. Only 17 *S*s out of the 126 (13.5%) could not be classified because they scored equally high in more than one system or not high enough in any.

The Sentence Completion Test of Integrative Complexity

The Sentence Completion Test (SC), developed by Schroder and Streufert (1962), is a projective instrument in which *S* is asked to

[4] If *S* scored within the same half-decile range on two systems and these scores were within the four highest half-deciles (percentiles 81-100), then ties were resolved as follows:

1. If one of the systems was System I, *S* was classified into System I.
2. If one of the systems was System IV and the other *not* System I, *S* was classified into System IV.

Such ties occurred in 12 cases. In no case, were System II and III tied. The above rules were adopted since the systems at the extremes (i.e., I and IV) appear to be most clearly delineated by the ITI; and since they are most likely to be used as experimental groups, resolution of ties in their favor is an aid to classification. One outcome of the use of these rules was to reduce the number of *S*s classified into Systems II and III relative to the other two systems. However, since these rules made it possible to classify more people, they were deemed practical.

respond to stems which imply interpersonal conflict, ambiguity, or the imposition of control. The SC used in this study contained six stems. Each completion is scored on three scales: degree of abstractness, degree of System II content, and degree of System III content. The first scale is a five-point scale while the latter two are four-point scales. Inter-scorer reliabilities from .70 to .98 have been reported by Schroder and Streufert (1962). For this sample, the following reliabilities (using Pearson's Product Moment Correlation) were obtained: .80 for the degree of abstactness scale, .80 for the System II content scale, and .51 for the System III content scale.[5]

Following essentially the procedure outlined by Schroder and Streufert (1962), *S*s were assigned to one of the four systems.[6] On this basis, 35 *S*s were classified as System I, 25 as System II, 19 as System III, and 27 as System IV. Twenty *S*s could not be classified (16%). In terms of reliability of classification (based on the profile of the three scale scores), 76 per cent of the rescored cases were classified as they had been originally. Moreover, about one-half of the disagreements involved cases originally designated as unclassifiable.

The Creativity Test Battery

The creativity battery consisted of the following tests:

1. *Gestalt Transformations* (Guilford, *et al.*, 1952). The subject is asked to choose an object from among five choices that has a part which will serve a specified purpose. None of the objects listed is ordinarily used or associated with the specified purpose which is to be served. The score was the total number of correct solutions. This test loads on the factor of "semantic redefinition" identified by Guilford and Merrifield (1960).

2. *Match Problems II* (Berger, Guilford, and Christensen, 1957). The subject is asked to indicate several different patterns of matches that can be removed to leave a certain number of squares or tri-

[5] The author is again grateful to Mr. Robert Nye for scoring the Sentence Completion Tests. He was totally unaware of the names and identifications of *S*s when scoring their sentences as was the author when he rescored half of the tests to obtain reliability estimates.

[6] System assignment based on SC scores differed from that suggested by Schroder and Streufert (1962) only insofar as ranges were defined on a relative rather than an absolute basis. Rather than saying, *a priori*, that the System IV range on the degree of abstractness scale is 4.0-5.0, System IV range on this scale was defined here by *S*s scoring in the top 20 per cent on this scale.

angles. The score was total number of correct solutions. This test loads on the factor of "figural adaptive flexibility" (Berger, *et al.*, 1957).

3. *Consequences* (Christensen, Merrifield, and Guilford, 1958). The subject is asked to list as many consequences as he can of certain hypothetical and unusual situations. Two scores were given: one for total number of relevant responses, and one for number of remote responses. Remote responses were defined as those more distant, temporally or geographically, than an immediate or obvious response. Inter-scorer agreement on judgments of obvious and remote on a randomly drawn sample of 220 responses amounted to 95 per cent.

This test loads on the factor of "originality" identified by Guilford and Merrifield (1960).

The four scores obtained from the three creativity measures were converted to Z scores. The criterion was a composite score labelled the *combined creativity score,* which was obtained by summing the four Z scores: Gestalt Transformations + Match Problems + Consequences: Total + Consequences: Remote. The combined creativity score is highly similar to the creativity criterion score used by Garwood (1964), the difference being that Garwood included three additional measures of creativity. A discussion of the validity of the creativity measures appears in Garwood (1961).

A measure of creative motivation, the Creativity Motive Questionnaire (Golann, 1962) was also included in the battery. This test presents the subject with a pair of tasks or situations, differing in the extent to which each will appeal to the creative motive, and he must indicate his preference. The score is the total number of "creative" tasks or situations chosen.

Results

The Prediction of Creative Performance

The means on the combined creativity score were compared by analysis of variance and Duncan's Multiple Range Test (Winer, 1962) for each of the four system groups, as classified separately by the ITI and SC. These means, and the outcomes of the analyses appear in Table 1.

As shown, means for the groups classified by the ITI were signi-

TABLE 1

Means and Analysis of Variance of Combined Creativity Scores

	Systems (means)				Analysis of Variance				
Test	I	II	III	IV	MS_B	df	MS_W	df	F
ITI (Objective)	38.55[a]	43.08	41.59[b]	44.40	170.9	3	39.9	105	4.28*
SC (Projective)	38.09[a]	42.12[c]	42.78	43.93	168.9	3	34.9	102	4.63*

* $p < .01$

[a] Significantly different from other 3 means by Duncan Range Test ($p < .01$)

[b] Significantly different from mean for System IV by Duncan Range Test ($p < .01$)

[c] Same as b, but $p < .05$

ficantly different ($F = 4.28, p < .01$[7]) as were means for the groups classified by the SC ($F = 4.63, p < .01$). In both cases, high F ratios were based primarily on differences between the creative performance of System I individuals versus all other systems. That is, using either instrument as a basis for classification, System I individuals performed less creatively than did individuals of the other three systems.

Among the other systems, only System IV and System III were significantly different from one another when the ITI was the basis for classification, while System IV and System II were significantly different from one another when the SC was the basis for classification. In the former case, II's exceeded III's and IV's exceeded II's, but these differences were not significant. Using the SC, III's exceeded II's and IV's exceeded III's, but these differences were not significant. Thus, creative performance for system groups classified by both instruments was highly similar, and in the predicted direction.

The Prediction of Creative Motivation

Mean Creativity Motivation Questionnaire (CMQ) scores and analyses for the system groups classified separately by the ITI and SC appear in Table 2. CMQ means of groups classified by the ITI did not differ ($F = 0.75$) while creative motivation of groups classified by the SC did differ ($F = 4.15, p < .01$). In the latter case, the high F ratio was based on differences between System I

[7] All probabilities listed in the text and the tables are two-tailed.

TABLE 2

Means and Analysis of Variance of Creativity Motive Questionnaire (CMQ) Scores

	Systems (means)				Analysis of Variance				
Test	I	II	III	IV	MS_B	df	MS_W	df	F
ITI (Objective)	19.42	21.15	20.41	21.63	25.1	3	33.7	105	0.75*
SC (Projective)	18.17[a]	20.92	20.95	23.63[a]	125.1	3	30.2	102	4.15**

* Not significant
** $p < .01$
[a] Significantly different from other 3 means by Duncan Range Test ($p < .01$)

and the other three systems, and differences between System IV and the other three systems. Of the six comparisons made, only the difference between System II and System III means was not significant. Clearly, system classification via the SC led to the identification of inter-system differences in creative motivation, while system classification via the ITI did not. These results were contrary to what was expected in the use of the ITI.

The Prediction of Intelligence

The four system groups, as classified separately by the ITI and SC, were also examined with regard to intelligence (Table 3). As shown, the system differences in creative performance and creative

TABLE 3

Means and Analysis of Variance of Intelligence (GCT) Scores

	Systems (means)				Analysis of Variance				
Test	I	II	III	IV	MS_B	df	MS_W	df	F
ITI (Objective)	59.50	60.50	59.71	60.62	7.9	3	18.3	98**	0.43*
SC (Projective)	58.45	60.72	59.94	61.28	35.2	3	16.8	95**	2.09*

* Not significant
** There are fewer degrees of freedom than in previous analyses because intelligence scores were not available for 9 *Ss*.

motivation could not be accounted for on the basis of intelligence, using either the ITI or the SC.

Direct Comparison between the Two Classification Instruments

Level of Abstractness as a Criterion. The Sentence Completion Test contains three scoring scales, one of which is a five-point level of abstractness scale. In using this scale to classify system-types, IV's are considered to be more abstract than III's, III's more abstract than II's, and II's more abstract than I's. The level of abstractness score was used as a criterion for evaluating system classification based on the ITI.

TABLE 4

Means and Analysis of Variance of Level of Abstractness Scores (taken from the Sentence Completion Test)

Test	Systems (means)				Analysis of Variance				
	I	II	III	IV	MS_B	df	MS_W	df	F
ITI (Objective)	1.92	1.99	2.17[b]	2.46[a]	1.62	3	0.47	105	3.45*

* $p < .05$
[a] Significantly different from other 3 means by Duncan Range Test ($p < .01$)
[b] Significantly different from mean for System I by Duncan Range Test ($p < .05$)

From Table 4 it can be seen that levels of abstractness for the groups classified by the ITI were different ($F = 2.45, p < .05$). This high F-ratio was based on significant differences between System IV and each of the other three systems, and between System III and System I. Although the System II mean was higher than that of System I, and the former was lower than that of System III, as expected, neither difference reached acceptable levels of significance.

ITI VS. SC: person-by-person classification. The most direct test of agreement between the two classification measures was to examine the extent of correspondence between classification made on the basis of one measure with that made on the basis of the other, with the individual as the unit. Of the 126 *Ss* tested, 94 were classified into one of the four systems by both measures. The conjoint classification of each of these *Ss* by the ITI and SC is shown in Table 5. This frequency data yielded a $X^2 = 38.02$ ($p < .01$), which was then converted to a contingency coefficient (Guilford, 1950) of .54 (out of a maximum C of .87), indicating a strong

TABLE 5

Cross-classification of Individuals by the Interpersonal Topical Inventory and the Sentence Completion Test

		Classification Based on the ITI I	II	III	IV	
Classification based on the SC	I	16	4	8	4	32
	II	4	12	2	3	21
	III	2	4	7	5	18
	IV	4	3	3	13	23
		26	23	20	25	94

relationship between the two classification instruments. An examination of Table 5 shows that inter-test agreement is at or beyond 50 per cent for all systems except System III where the extent of disagreement is high. Here *S*s classified as System III by the ITI are as likely to be classified as System I by the SC as they are to be classified as System III.

Discussion

Using the criterion of creative performance, System I *S*s were found to do significantly worse than *S*s from any of the other three systems regardless of whether the ITI or the SC was used as the system classification instrument. In general, I's were found to be least creative, and IV's most creative with II's and III's in-between. The results leave the location of Systems II and III with respect to each other and to the other two systems somewhat ambiguous. Based on the ITI as the classification instrument, III's are substantially less creative than IV's. Based on the SC as the classification instrument, II's are substantially less creative than IV's. The latter finding is more in line with expectations.

The ambiguity can be somewhat resolved by examining the area of disagreement between the two classification instruments. Table 5 shows that many individuals classified into System III by the ITI are classified into System I by the SC. If the ITI is weak in this area of classification, then these individuals are, in fact, misclassified I's. Since I's are the least creative of all system types, if they are being erroneously included in the ITI System III group then

the mean for this group is being artifactually lowered. This would account for the transposition of II's and III's on creative performance when the ITI is the classification instrument.

This conjecture was examined more systematically by comparing the creative performance scores for the seven *S*s who were classified as System III by both instruments with the eight *S*s who were classified as System III by the ITI and System I by the SC. For the former seven, five of whom had creativity scores of 40 or above, a mean creativity score of 42.7 was obtained. For the latter eight, three of whom had creativity scores of 40 or above, a mean creativity score of 40.0 was obtained. While the number of *S*s involved are not sufficient for a meaningful statistical test, it appears plausible that some System I individuals were inaccurately classified as System III by the ITI, producing a spuriously low mean creativity score for System III.

It was concluded that there is a strong relationship between creative performance and level of integrative complexity based on the analysis of mean combined creativity scores for system groups identified by the SC. It was also concluded that the ITI is a useful research instrument and reasonable substitute for the SC for identifying Systems I, II, and IV but is inadequate in its present form for identifying System III. This conclusion was based on the predictability of creative performance when the ITI was used as the classification instrument and the extent of agreement between the two system-measures. However, the adequacy of the SC for measuring System III is also questionable owing to the low inter-scorer reliability obtained for the System III scale of the SC.

Inter-system differences in creative performance were not explicable on the basis of intelligence since systems did not differ significantly in intelligence using either classification instrument. Intelligence effects were somewhat higher when the SC was the classification instrument but this is probably due to the fact that verbal ability plays a bigger role in writing an original response than choosing from among given responses.

The SC as a classification instrument led to predictable differences between systems on creative motivation while the ITI did not. It may be that the SC, as a projective, open-ended instrument, is more sensitive to individual motivation than is the ITI. However, the "price" paid for this heightened sensitivity is, in many cases,

excessive. The SC is difficult to score, and is especially limiting when rapid scoring is necessary.

The ITI was designed to be an adequate, easily scored measure of the systems of integrative complexity, primarily for use as a research tool. Based on the findings of this study, it was concluded that the ITI significantly discriminated between systems, but that further efforts were needed to improve its capacity to measure System III. Such efforts should be based on criteria other than correspondence to the SC.

Summary

This study attempted to evaluate an objective measure of level of integrative complexity, the Interpersonal Topical Inventory (ITI). Evaluation was based on the extent to which the objective ITI could predict creative performance and creative motivation as well as scores on a projective instrument presently in use as a measure of integrative complexity, the Sentence Completion Test (SC). A further source of evaluative data was the correspondence between the new classification instrument and the projective SC.

The concept of integrative complexity represents the extent to which the individual perceives his world and those in it in a highly differentiated and integrated manner. The greater the individual's capacity for differentiation and integration, the more likely he is to produce creative responses. On this basis, creative performance was deemed an evaluative criterion for measures of integrative complexity.

Data from 126 Naval enlistees showed that classification by either the ITI or the SC led to similarly accurate predictions of creative performance, as measured by a battery of creativity tests. This was not the case for creative motivation, where accurate predictions followed only from SC classification. However, the degree of correspondence in classification of individuals by the two instruments was high at all levels of integrative complexity except one. In no case were differences in integrative complexity accountable for on the basis of intelligence.

It was concluded that the objective ITI is a potential substitute for the projective SC in large scale surveying of individuals for research.

REFERENCES

Berger, R. M., Guilford, J. P., and Christensen, P. R. A. Factor-Analytic Study of Planning Abilities. *Psychological Monograph*, 1957, 71, No. 6 (Whole No. 435).

Christensen, P. R., Merrifield, P. R., and Guilford, J. P. *Consequences: Manual for Administration, Scoring, and Interpretations.* Beverly Hills, California: Sheridan Supply, 1958.

Driver, M. J. Conceptual Structure and Group Processes in an Interaction Simulation. I. The Perception of Simulated Nations. Princeton, New Jersey: ONR Technical Report #9, Nonr 1858 (12), 1962.

Garwood, Dorothy S. Some Personality Factors Related to Creativity in Young Scientists. Unpublished doctoral dissertation, Claremont College, 1961.

Garwood, Dorothy S. Personality Factors Related to Creativity in Young Scientists. *Journal of Abnormal and Social Psychology*, 1964, 68, 413–419.

Golann, S. E. The Creativity Motive. *Journal of Personality*, 1962, 30, 588–600.

Guilford, J. P. *Fundamental Statistics in Psychology and Education* (2nd. ed.). New York: McGraw-Hill, 1950.

Guilford, J. P. and Merrifield, P. R. The Structure of Intellect Model: Its Uses and Implications. *University of Southern California Psychology Laboratory Report*, 1960, No. 24.

Guilford, J. P., Wilson, R. C., and Christensen, P. R. A Factor-Analytic Study of Creative Thinking: II. Adminstration of tests and analysis of Results. University of Southern California *Psychology Laboratory Report*, 1952, No. 4.

Harvey, O. J., Hunt, D. E., and Schroder, H. M. *Conceptual Systems and Personality Organization.* New York: Wiley, 1961.

Lawrence, E. A. An Investigation of Some Relationships between Personality Structure and Group Functioning. Unpublished manuscript, Princeton University, 1962.

Schroder, H. M. and Streufert, S. The Measurement of Four Systems of Personality Structure Varying in Level of Abstractness. (Sentence Completion method.) Princeton University: ONR Technical Report #11, Nonr 1858 (12), 1962.

Streufert, S. Attitude Generalization in Social Triads as a Function of Personality Structure and Availability of Social Support. Princeton University: ONR Technical Report #10, Nonr 1858 (12), 1962.

Suedfeld, P. Attitude Manipulation in Restricted Environments: I. Conceptual Structure and Response to Propaganda. *Journal of Abnormal and Social Psychology*, 1964, 68, 242–247.

Tuckman, B. W. Personality Structure, Group Composition, and Group Functioning. *Sociometry*, 1964, 27, 469–487.

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

A SHORT TEST OF ONE'S EDUCATIONAL PHILOSOPHY

R. L. CURRAN, I. J. GORDON and J. F. DOYLE
University of Florida

This is a time in educational history when knowledge and knowing are again being heavily stressed as goals for education, and when new curricula are being developed to achieve these goals. Simultaneously, research on the "anatomy of teaching" and the "logic of teaching" is being conducted to examine what the teacher does when he teaches, without regard to the what and the why. It is important, however, to examine a teacher's philosophy of education, defined here as a logically inter-related set of concepts of reality, knowledge and values. This philosophy must be understood in conjunction with the analytical study of teaching in order to gain insight into just what the teacher views to be the goals of education. This study, therefore, was undertaken to develop a short, reliable and valid instrument to measure the ontological, epistemological and axiological dimensions of a teacher's philosophy of education.

An inspection of research publications and the professional literature since 1952 reveals the absence of any widely used instruments for measuring a teacher's philosophy of education. Among the few efforts to develop a measurement instrument of this type is a two-philosophies (empirical-rationalistic) Q-sort instrument called the GNC (Gowin, Newsome and Chandler, 1961). E. V. Sayers developed a paper-pencil test of consistent agreement with the educational philosophy of experimentality and proceeded as far as item analyses and reliability tests on education student populations in Hawaii and New York City.

Ryans (1961), as part of the Teacher Characteristics Study, de-

veloped an estimate of conservative educational viewpoint. This score measured the dimension called conservative, learning-centered—liberal, child-centered. Ryans reports that scores on this twenty-item forced-choice instrument revealed differences between teachers related to grade level and subject matter taught, years of experience, and age. Although this is the most thorough and systematic work to date, it is not really a measure of a philosophy of education but rather of a psychological dimension of attitude toward pupils and those aspects of teacher behavior which relate to inter-personal relationships, teacher's organization (poised-excitable, disorganized-systematic, etc.) and mode of presentation (dull-stimulating, sterotyped-original).

Kerlinger (1961) has developed scales which measure educational "progressivism" and "traditionalism," but not the identity and logical consistency of a person's ontological, epistemological and axiological conceptions. His Likert-type scales grew out of "Q" sort beginnings. Each scale consists of 10 "progressive" and 10 "traditional" items. He reports: "This study must still be considered exploratory. . . . More work needs to be done with different items and different samples and longer scales" (Kerlinger, 1961, p. 284).

Procedure

With its one hundred items devised with the help of scholars of philosophy the GNC was easily the most extensive and authoritative source of items and thus the obvious resource with which to begin. The GNC was transformed from a Q-sort instrument into an ordinal attitude scale and administered to undergraduate and graduate classes in the philosophy of education at the University of Florida. Upon item analysis, 40 of these 100 GNC items yielded significant discriminatory power to measure the degree and consistency to which a person's conception of education is experimental or rationalistic in the three areas of ontology, epistemology and axiology.

These forty items were then successively combined with those items from the work of Sayers, Ryans, Kerlinger and Oliver (1953) which were felt to be "philosophic." In addition, a set of epistemological items was included from the work by a faculty committee charged with the development of a list of concepts which were

thought to be important for graduates of the college to hold. Further item analyses were made on the new instruments thus evolved.

These successive item analyses yielded a final pool of fifty items which had, over the several test administrations with graduate and undergraduate University of Florida classes, maintained statistically significant discriminatory power. The task then shifted to selecting from these fifty items a short schedule of items which would reliably and validly measure groups on the continuum of a conceptual philosophy of education that ranged from most rationalistic to most experimentalistic.

Sample for Final Analysis

Table 1 presents the number and arrangement of the groups whose test scores were used to provide the data for the final item analysis and reliability study. Two administrations of the test were given, a month apart, since reliability was to be of the test-retest model. It was anticipated that this design would provide three distinct groups: group A being those who were present for both administrations of the test; group B those who were present for the first administration but not the second; and group C those who were present for the second administration but not the first. As it turned out, group A yielded a sample of 292 persons, while groups B and C numbered 62 and 58 students, respectively.

TABLE 1

Number and Arrangement of Groups Used in Item Analysis and Reliability Test Administrations

TEST ADMINISTRATION					
TIME I			TIME II		
GROUP		N	GROUP		N
A_i	Total Group	292	A_{ii}	Total Group	292
	Graduates	78		Graduates	78
	Seniors	117		Seniors	117
	Sophomores	97		Sophomores	97
B	Total Group	62	C	Total Group	58
	Graduates	14		Graduates	19
	Seniors	23		Seniors	28
	Sophomores	25		Sophomores	11
TIME I TOTAL		354	TIME II TOTAL		350

The test results from group A (Time I) were used to make the

item analysis, and these same test results taken in conjunction with the Time II test administration with Group A supplied the data upon which the test-retest reliability estimate was made. In order to cross validate the results of the item analysis performed with the group A test scores, groups B and C were combined into one group of 120 subjects and the analysis was repeated. No other use had been made of the results from group B and C, thereby making them a legitimate and time-saving second sampling suitable for the purpose of cross validating the initial item analysis.

The problem of empirical validity was approached through selecting a sample that would be representative of populations which had had different amounts of exposure to the University of Florida's program of teacher education. With this criterion in mind, three sub-groups were established: sophomores, seniors, and graduate students. The sophomore group consisted of those taking their first courses in education and were, for the most part, those enrolled in courses in human growth and development. The senior group were those who were just finishing their undergraduate programs, but who had not as yet begun their teaching internships. Finally, the graduate group comprised students who had entered a graduate program directly from undergraduate study as well as many who had returned to the University after an indeterminate number of years in teaching or other fields.

It was assumed that the University of Florida College of Education faculty generally subscribes to an experimental conception (as expressed by Dewey, Mead, Kilpatrick, Bode, Kelley, Stanley and the like). Therefore, it was further assumed that a division of the student body into sophomore, senior and graduate level exposure to this position would yield a criterion population with the graduates tending to express the more experimental and the sophomores the more rationalistic philosophy of education.

Data Analysis

Two measures of item discriminatory power were used, each of them based on the upper and lower 27 per cent of the distribution. The *t*-score discrimination index is that suggested by Edwards (1957) and the PHI index is that recommended by Guilford (1954). The *t*-score index is suitable for use in those situations in

which items are scored on Likert-type scales, while the PHI index requires the more traditional correct-incorrect type of scoring.

Subjects scored items on a four point Likert-type scale, the category ratings being the following: Agree Strongly, Agree, Disagree, Disagree Strongly. Responses on the non-experimental items were scored 1, 2, 3, and 4, respectively. For the experimental items the scoring weights were simply reversed. Computation of the *t*-score index therefore was based on fully weighted scores, while computation of the PHI index required that scale values 1 and 2 be compressed into the single value 0, and scale values 3 and 4 into the value 1.

Item difficulty level (D) was computed on the basis of the compressed correct-incorrect (in this case, experimental–nonexperimental) scoring, using the total number of subjects.

In selecting items for inclusion in the final form of the instrument, consideration was given to the three indices reported. Each index, of course, was interpreted in the light of its own criterion. In the case of the *t*-score index, the suggested procedure is to select the required number of items in the order of the magnitude of *t*, taking a *t* of 1.75 as a minimally acceptable lower limit (Edwards 1957, p. 152). Following Guilford's procedures for estimating the significance of PHI with samples of different sizes, the required magnitude of indices in this study were: for the initial analysis, .11 at the .05 level of significance, and .14 at the .01 level; for the validation analysis, .18 at the .05 level, and .23 at the .01 level of significance (Guilford, 1954, p. 432). With respect to item difficulty it was decided that for the purposes of this study, no particular concern would be expressed for any item remaining within a range bounded by .20 and .80.

Results

For the most part, these criteria were met in the final item analysis and sustained in the cross validation analysis for the twenty-five items selected for inclusion in the final form of the instrument.

As a result of the final item analysis, the following twenty-five items were selected as the most usable in a short test which would measure a subject's predisposition to express a philosophy of education that could be termed experimentalistic.

1. In this period of rapid change, it is highly important that edu-

cation be charged with the task of preserving intact the long established and enduring educational aims and social objectives.

2. The true view of education is so arranging learning that the child gradually builds up a storehouse of knowledge that he can use in the future.

3. In assessing what man knows, there are no absolutes, only tentative conclusions based on the current accumulation of human experiences.

4. Required reading of literary works, even though it may bring an unfavorable attitude toward literature, is necessary in a sound educational program.

5. To learn means to devise a way of acting in a situation for which old ways are inadequate.

6. In the interest of social stability, the youth of this generation must be brought into conformity with the enduring beliefs and institutions of our national heritage.

7. Learning is a process of mastering objective knowledge and developing skills by drill, trial and error, memorization, and logical deduction.

8. The teacher must indoctrinate her students with correct moral principles in order to bring about their healthy moral development.

9. Moral education is the continuous criticism and reconstruction of ideals and values.

10. The traditional moral standards of our culture should not just be accepted; they should be examined and tested in solving the present problems of students.

11. The backbone of the school curriculum is subject matter; activities are useful mainly to facilitate the learning of subject matter.

12. A teacher may properly teach that some laws are unchanging and certain in their essential nature.

13. Moral learning is experimental; the child should be taught to test alternatives before accepting any of them.

14. Minimum standards of achievement, in the form of requirements to be met equally by all students, must be demanded at every level of education.

15. Existing knowledge is tentative and is subject to revision in the light of new facts.

16. A knowledge of history is worthwhile in itself because it embraces the accumulated wisdom of our ancestors.

17. An activity to be educationally valuable should train reasoning and memory in general.

18. The teacher is a channel of communication, transmitting knowledge from those who know to those who do not know.

19. The best preparation for the future is a thorough knowledge of the past.

20. The curriculum should contain an orderly arrangement of subjects that represent the best of our cultural heritage.

21. Child life is not a period of preparation, but has its own inherent value.

22. The aim of instruction is mastery of knowledge.

23. There is no reality beyond that knowable through human experience.

24. Learning is essentially a process of increasing one's store of information about the various fields of knowledge.

25. Only that should be accepted as true which meets the test of experience.

Table 2 presents the item analysis results for both the final analysis and the cross validation analysis. In addition, each item is identified as either an E (experimental statement) or an R (rationalistic statement), classification dependent solely on the item's content value. Support for the validity of these classifications rests, of course, on the fact that all but items No. 3, No. 15, No. 21, No. 23 and No. 25, were identified as experimental or rationalistic in previous instruments or judgments. These five items were from the original set developed by the faculty committee.

When subjected to the cross-validation analysis, one item, No. 25, fell below both the *t*-score and the PHI criteria for admissability and is therefore not recommended for future use. Item No. 15 did not reach significance on the PHI index. Also, this item, together with item Nos. 1 and 3, had a difficulty level which was outside the criterial range. It is recommended, however, that these three items be retained. The deviation of item No. 1 from the criterion of .80 was slight and occurred only in the initial analysis. Item No. 3, despite a high difficulty level, still had significant discriminatory power. It is hoped that further administrations of the

TABLE 2

Item Analysis Results Reported in Terms of Discrimination and Difficulty Level Indices

Item Number	Item[a] Source	Content Classification: Rationalistic or Experimental	Final Analysis (N = 292) Discrimination Difficulty			Cross Validation Analysis (N = 120) Discrimination Difficulty		
			t-score	PHI	D	t-score	PHI	D
1.	GNC	R	6.15	.35	.82	6.22	.43	.80
2.	K	R	5.98	.49	.42	4.58	.60	.49
3.	UF	E	5.42	.19	.90	5.16	.18	.81
4.	GNC	R	7.85	.57	.58	5.79	.80	.58
5.	S	E	2.53	.25	.63	3.29	.20	.58
6.	GNC	R	4.35	.40	.66	6.61	.60	.69
7.	GNC	R	8.34	.58	.51	4.82	.57	.50
8.	GNC	R	8.24	.49	.72	5.88	.67	.68
9.	S	E	3.15	.28	.70	2.20	.18	.66
10.	K	E	5.16	.38	.70	4.55	.47	.75
11.	K	R	6.83	.49	.55	4.12	.40	.57
12.	GNC	R	6.14	.47	.39	6.04	.47	.42
13.	K	E	5.98	.44	.37	2.78	.43	.39
14.	GNC	R	8.31	.64	.60	3.12	.50	.54
15.	UF	E	4.68	.03	.98	3.36	.10	.97
16.	GNC	R	6.10	.28	.70	5.16	.60	.40
17.	R	R	7.81	.56	.40	4.61	.43	.37
18.	GNC	R	9.20	.67	.39	3.26	.53	.41
19.	GNC	R	2.53	.44	.60	4.20	.43	.57
20.	K	R	4.21	.31	.37	4.22	.40	.46
21.	UF	E	7.96	.40	.72	3.73	.40	.69
22.	GNC	R	8.24	.51	.70	3.70	.47	.70
23.	UF	E	5.29	.40	.43	2.47	.20	.34
24.	K	R	9.55	.75	.44	4.69	.50	.47
25.	UF[b]	E	2.78	.31	.29	.15	−.10	.45

Significance of PHI: for $N = 292$: $p < .05$, .11; $p < .01$, .14.
for $N = 120$: $p < .05$, .18; $p < .01$, .23.

[a] GNC—Gowin, Newsome, and Chandler
K—Kerlinger
UF—Faculty Committee, College of Education, University of Florida
S—Sayers
R—Ryans

[b] Not recommended for inclusion in future use of the instrument.

instrument with groups less experimental than those at the College of Education, University of Florida, will reduce the difficulty level of Item No. 15 and possibly also increase its discriminatory power. In the meantime, it serves the purpose of keeping the number of E items in the instrument from falling below a third of the total number of items. Even if future administrations of the instrument do not vindicate the inclusion of item No. 15 on the basis of in-

creased discriminatory power, it should still be kept in order to hold the type of items in balance and possibly avoid response "sets." Under these circumstances, the item should, however, be disregarded in computing a subject's score on the instrument.

TABLE 3

Means, Standard Deviations, and Reliability r's for Final Analysis

TEST ADMINISTRATION										
TIME I					TIME II					
	Group	N	MEAN	S.D.		Group	N	MEAN	S.D.	
A_I	Total Group	292	67.49[a]	8.02	A_{II}	Total Group	120	67.06[a]	8.19	.82
	Graduates	78	71.82*	9.22		Graduates	78	72.12*	9.16	.83
	Seniors	117	67.56*	6.38		Seniors	117	67.44*	6.78	.72
	Sophomores	97	63.70*	6.73		Sophomores	97	62.77*	6.32	.80

*$p < 01$

[a] Maximum possible score = 100, minimum = 25.

As Table 3 shows, the test-retest reliability correlation coefficient for the total group of 292 students was .82. When the subgroups composing the total sample were examined, comparable correlations (.83 and .80) were found for the graduate students and the sophomores. The seniors, however, had a reliability correlation of only .72. Table 3 also reports the subgroup mean scores, together with their standard deviations for both the Time I and Time II test administrations. The instrument had been scored in such a way that the more experimental the educational philosophy of the subject the higher the score he would receive. The graduate students were significantly more experimentalistic than the sophomores. That this could be taken as partial confirmation of the validity of the instrument resided in the fact, already mentioned, that this should be the expected distribution in the light of the type of professional development offered at the University of Florida's College of Education.

In order to establish the stability of these differences found among the graduate, senior and sophomore groups, a replication was performed with another sample from these populations. The results are found in Table 4. In this table, since all the seniors available for testing were currently interning, this group is referred to as interns.

A change was made in the method of scoring the instrument. The

TABLE 4

Test Validity Reported on Basis of Replication of Significant Sub-Group Mean Differences

Group	N	Mean	S.D.
1. Graduates[a]	42	63.50*	11.83
2. Interns	211	57.36*	10.66
3. Sophomores[a]	50	46.98*	7.79

* Different from other two groups, $p < .001$

[a] Point-biserial r between total score and group membership was .63 ($p < .01$)

scoring scale was expanded from a four-point scoring band of Agree Strongly, Agree, Disagree and Disagree Strongly to one that included a middle category of No Opinion. The weighting of responses was also changed from a 1 through 4 range to a 0 through 4 range. As Table 4 indicates, the subgroup means were again found to differ significantly and in the direction expected. Validity was also demonstrated by finding that a point biserial correlation of .63 existed between total test scores and a student being in either the graduate (most experimental) or sophomore (least experimental) group. If, as it seems reasonable to assume, the student body at the University of Florida College of Education becomes more experimental and less rationalistic in stated philosophic outlook the longer a student has been in attendance, then the instrument presented here appears to be a valid, as well as short and reliable, measure of that dimension of an individual's philosophy of education.

In constructing this instrument it was discovered that the population sampled was skewed in the direction of the experimentalist position. Despite the paucity of subjects possessing a rationalist philosophy of education, the items were able to yield satisfactory discriminatory power. It seems reasonable to conjecture that both item discrimination and test validity coefficients would be strengthened if the test were now to be administered to a criterion sample which included a larger representation of rationalist subjects. Verification of this expectation would appear to be the next logical step in any further research on this instrument.

REFERENCES

1. Edwards, A. L. *Techniques of Attitude Scale Construction*. New York: Appleton-Century Crofts, Inc., 1957.

2. Gowin, D. B. Newsome, G. L., and Chandler, A. K. Scale to Study Logical Consistency of Ideas about Education. *Journal of Psychology*, 1961, 51, 443–455.
3. Guilford, J. P. *Psychometric Methods*, New York: McGraw-Hill, 1954.
4. Kerlinger, F. Factor Invariance in the Measurement of Attitudes Toward Education. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 2, 275–285.
5. Oliver, W. A. Teachers' Educational Beliefs versus Their Classroom Practices. *Journal of Educational Research*, 1953, 47, 47–55.
6. Ryans, D. Inventory Estimated Teacher Characteristics as Covariants of Observer Assessed Pupil Behavior. *Journal of Educational Psychology*, 1961, 52, 91–97.
7. Sayers, E. V. Personal Communication with Robert L. Curran.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

SOME DIMENSIONS OF MEANING OF THE CONCEPT TELEVISED INSTRUCTION[1]

FRED W. OHNMACHT[2]
University of Maine

The flexibility, efficiency and economy of the Semantic Differential has led to its widespread application in the measurement of attitudes towards a variety of stimuli. Recent studies by Greenberg (1963), Janes (1964), Kraus (1962) and Westley and Jacobson (1963) are among the many which testify to the popularity of the technique. Typically, investigators have selected scales which represent the Evaluative, Potency, and Activity factors as identified by the work of Osgood (1957) and his associates. Responses to the scales are compared individually and values for the three factors are derived through the summation of the scales representing each one. The summed evaluative scales are thought to represent attitude toward the stimulus rated.

Now the fact that Evaluation, Potency, and Activity factors typically turn up when matrices of intercorrelations of scales are factored is a most interesting finding. However, this in no way suggests that scales loading on these factors when they are first summed across a large number of concepts are appropriately used for the estimation of the aforementioned factors when *specific* concepts are the point at issue. This point does not appear to be sufficiently appreciated by a number of investigators.

Osgood (1957) has pointed out that when responses to a Semantic Differential for a single concept are factor analyzed a good deal of concept-scale interaction may take place. A recent study by

---

[1] The research reported here was supported by the College of Education Team Teaching Project which is funded by the Ford Foundation.
[2] Now at the University of Georgia.

Husek and Wittrock (1962) demonstrated the same point when it was found that scales utilized as estimators of the Potency and Activity factors aligned themselves with other evaluative scales when the concept "School Teachers" was rated by means of a form of the Semantic Differential. Hartman (1963), in his recent review, suggests that users of the Semantic Differential would be well advised to perform their own factor analyses prior to using the technique. This suggests that investigators should be wary of measuring the meaning of specific concepts with scales identified through analyses which have collapsed data across concepts.

Purpose

The purpose of the present study was to explore the nature of the meaning of the concept Televised Instruction. The analysis undertaken clarifies the nature of the concept-scale interaction and suggests scales which can properly be used in assessing attitudes with respect to the concept under consideration.

Methods and Procedures

A group of 105 students enrolled in a graduate seminar at the University of Maine rated the concept Televised Instruction with a form of the Semantic Differential consisting of 26 scales. The scales were selected on the basis of presumed relevance to the concept in question. A number of scales were included to insure marker representation for the Evaluative, Activity, and Potency factors if they characterized dimensions underlying the data.

The scores on the 26 scales were intercorrelated and the resulting matrix subjected to a principal-components analysis. Unities were employed in the main diagonal and all components retained for rotation whose latent roots exceeded one, as suggested by Kaiser (1960). The resulting principal-component structure was subjected to varimax rotation (Kaiser, 1958). The scale loadings on the resulting orthogonal factor structure were interpreted through an examination of scales having a loading of .40 or higher on a given factor. The location of such marker scales as good-bad (evaluation), strong-weak (potency) and active-passive (activity) also aided in the interpretation of the rotated components.

Findings

Seven principal components extracted from the 26 $\times$ 26 matrix of scale intercorrelations had latent roots greater than unity. These seven components were rotated to the varimax criterion. Table 1 presents the loadings of the 26 scales on the seven rotated reference axes. In order to simplify the table, only loadings of .40 or higher are shown.

An examination of the hyperplane counts ($\pm$.15) for the seven rotated factors discloses a reasonably clear simple structure. Factor I, however, appears to be a general factor with but five scales having a loading within .00 $\pm$ .15.

Factor I–General Evaluation

This factor is clearly a general evaluative factor which accounts for approximately twice as much variance as any other factor. Scales such as good-bad (.76), meaningful-meaningless (.81) and positive-negative (.57) mark the factor as evaluative in nature. The general factor subsumes several "modes" of evaluation since the scales colorful-colorless (.69) and interesting-boring (.70) have their major loadings on this factor. These scales have been identified by Osgood (1957) as loading on a Receptivity factor which represents a mode of evaluation independent of the more commonly found Evaluation factor. Perhaps the most striking loading on Factor I is that of the weak-strong scale (—.78). This scale is usually thought of as a marker for the Potency factor. Osgood (1957) has suggested that the greater emotionality involved in a concept, the greater the tendency for scales representing evaluation and other dimensions found in more general analyses to converge. The rotational affects noted with respect to Factor I suggests that this is the case for the concept Televised Instruction and reinforces the notion that the semantic differential must be adapted to the requirements of a research problem, preferably by a preliminary analysis of concept-scale interaction when attitudes toward particular stimuli are at issue.

Factor II–Relation of TV Instruction to Pupil

Five scales have a rotated loading in excess of .40 on this dimension. The scales warm-cool and passive-active have their ma-

TABLE 1

Rotated Factor Loadings of Scores on Twenty-Six Semantic Differential Scales (Concept: Televised Instruction)

Scale	Factor I	II	III	IV	V	VI	VII	h_2
1. valuable-worthless	.46					.49		.65
2. heavy-light							.72	.59
3. unfair-fair			.58					.56
4. small-large			.65					.57
5. fast-slow						.74		.70
6. unimportant-important	−.60							.64
7. colorful-colorless	.69							.58
8. interesting-boring	.70							.74
9. mild-severe					.83			.72
10. unpleasant-pleasant	−.47		.58					.76
11. positive-negative	.57							.56
12. weak-strong	−.78							.74
13. complex-simple				.86				.79
14. changeable-stable						.57	.40	.59
15. disreputable-reputable			.78					.67
16. meaningful-meaningless	.81							.76
17. good-bad	.76							.72
18. impersonal-personal		.56		−.42				.59
19. masculine-feminine					.53		.52	.74
20. passive-active		.68						.72
21. deep-shallow	.41				.64			.74
22. successful-unsuccessful	.66							.59
23. warm-cool		−.74						.70
24. excitable-calm	.41							.58
25. constrained-free		.55					.40	.61
26. sensitive-insensitive	.40	−.62						.58
Hyperplane Count ±.15	5	14	10	15	18	12	15	
% of Total Variance	20.6	9.5	10.5	5.9	6.7	7.0	5.9	
% of Common Variance	31.2	14.3	15.9	8.9	10.1	10.6	8.9	

jor loadings on this factor. The configuration of loadings indicate that Factor II is an Activity factor. The loadings of the sensitive-insensitive (−.62), constrained-free (.55) and impersonal-personal (.56) further suggest that this factor represents a dimension of meaning similar to one identified by Osgood (1957) as Oriented Activity.

Factor III–Personal Feeling about TV Instruction

This factor is most clearly represented by the reputable-disreputable (.78) scale. Other scales having loadings in excess of .40 are small-large (.65), unfair-fair (.58) and unpleasant-pleasant (.58). The interpretation of this factor does not seem as clear cut as for Factor I and II.

Factor IV–Complexity (Specific)

Factor IV is highly specific to the complex-simple (.86) scale with the impersonal-personal (–.42) being the only other scale with a loading in excess of .40. The lack of other scales with reasonably high loadings on this factor make its identification tenuous at best.

Factor V–Severity

Factor V is characterized by the scales mild-severe (.83), deep-shallow (.64) and masculine-feminine (.53). These scales often are loaded on a Potency factor but in the present instance scales loading on Factor VII also suggests a Potency factor. It would appear that the Potency factor usually found has been broken up in the present analysis into two separate factors with some scales (strong-weak and sensitive-insensitive for example) usually found to load on the Potency factor rotating to the Evaluation and Activity dimensions.

Factor VI–Pacing

This factor is characterized by the scales fast-slow (.74), changeable-stable (.57) and valuable-worthless (.49). The first two scales mentioned suggest that the Activity dimension has been separated into two independent parts (Factor II and VI). The valuable-worthless scale also loads on the Evaluative factor and suggests that subjects conceive of televised instruction which is quickly paced and changeable as being "good."

Factor VII–Dynamism

This factor, along with Factor IV, accounts for the smallest amount of variance. The scales heavy-light, masculine-feminine, excitable-calm, and changeable-stable have loadings in excess of .40 on this factor. It is interesting to note the position of the mascu-

line-feminine scale in the semantic space, especially with respect to Factors VII and V. On the one hand (Factor V), feminine tends to go with severe and shallow, while on the other hand (Factor VII), feminine goes with light, calm and stable. This suggests some ambivalence in the allocation of the masculine-feminine scale in the semantic space.

Application of Evaluative Scales

Scales having a loading of .60 or higher on the General Evaluation Factor (Factor I) were used a number of times to assess college students attitudes towards video taped televised lessons which they have just finished viewing. The seven scales thus used are scaled from one to seven and an attitude or evaluation score is obtained by summing the weights for the individual items to obtain an attitude score. In seven such applications the mean values obtained have ranged from 23.58 to 41.56, indicating that the college students involved do discriminate among various televised lessons in terms of the General Evaluative dimension. Total-score scale reliabilities for the seven administrations have ranged from .91 to .96 in terms of Hoyt analysis of variance estimates.

Discussion and Conclusions

The findings of this study indicate that Evaluation is a major dimension of the meaning of the concept televised instruction. The presence of such a dimension is consistent with a growing body of work with the semantic differential. Factors representing the Activity and Potency dimensions are not as clearly delineated. In part, this may be a function of the pool of 26 scales used. Several scales usually associated with dimensions of meaning other than Evaluation had their major loadings on the Evaluation factor. This indicates that concept-scale interaction occurs when the concept televised instruction is rated using the semantic differential. This is especially clear when it is observed that the scale weak-strong, which is a standard marker for the Potency factor, has a loading of —.78 on the Evaluative factor.

In the present case, an investigator who chose three or four scales each to represent the Evaluation, Potency, and Activity factors from the general analyses of Osgood would run the risk of dealing in numerology rather than establishing scores representing

the three dimensions of interest. The inclusion of the scale weak-strong to represent Potency would be basically in error and it is doubtful that it would be included as an estimate of Evaluation as the present study indicates it should be. The extent to which such concept-scale interaction occurs when other concepts are rated is typically unknown but is subject to empirical test.

The present study provides a pool of items with known factorial composition for the study of attitudes towards televised instruction. In addition this study reemphasizes the need for investigators to perform their own factor analyses when the nature of concept-scale interaction with respect to a given study are unknown. The semantic differential is not a definite set of items to be used with any and all concepts and must be adapted to the requirements of each research problem.

REFERENCES

Greenberg, Bradley S. Operation Abolition vs. Operation Correction. *AV Communication Review*, 1963, 11, 40–46.

Hartman, Frank R. A Behavioristic Approach to Communication. *AV Communication Review*, 1963, 11, 155–190.

Husek, T. R. and Wittrock, M. C. The Dimensions of Attitudes Towards Teachers as Measured by the Semantic Differential. *Journal of Educational Psychology*, 1962, 53, 209–213.

Janes, Robert W. Preexisting Attitudes of College Students to Instructional Television. *AV Communication Review*, 1964, 12, 325–326.

Kaiser, Henry F. The Application of Electronic Computers to Factor Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 141–151.

Kaiser, Henry F. *Comments on Communalities and the Number of Factors.* Paper Read at an Informal Conference at Washington University, St. Louis, May 14, 1960.

Kaiser, Henry F. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 1958, 23, 187–200.

Kraus, Sidney. Modifying Prejudice: Attitude Changes as a Function of the Race of the Communicator. *AV Communication Review*, 1962, 10, 14–22.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning.* Urbana: University of Illinois Press, 1957.

Westley, Bruce H. and Jacobson, Harvey K. Instructional Television and Student Attitudes Towards Teacher, Course and Medium. *AV Communication Review*, 1963, 11, 47–60.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

VALIDITY STUDIES SECTION

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

ANNOUNCEMENT REGARDING VALIDITY STUDIES

THE VALIDITY STUDIES SECTION is published *twice a year*, once in the Summer issue and again in the Winter issue, for which the closing dates for receiving manuscripts are February first and August first, respectively. Although articles between two and eight printed pages are usually preferred, an occasional exception is made to publish articles of somewhat greater length.

Considerable flexibility exists concerning format as can be seen from a study of recently published articles. However, the model presented in the Spring, 1953, issue of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT still represents a close approximation to what is customarily published. The prospective contributor is encouraged to read the original announcement.

In order that the usual number of articles of other types may not be reduced, it is necessary to enlarge the journal and to charge the authors for most of the publishing costs. For a running page of printed text the cost is fifteen dollars per page with extra charges for tables and complex material. Each author receives 100 free reprints.

Manuscripts should be sent to:

Dr. William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

COMPARISON OF CRITERION CLUSTERS OBTAINED BY ANALYZING THE HOMOGENEITY OF A SET OF REGRESSION EQUATIONS AND THE MATRIX OF INTERCORRELATIONS

JOHN P. CAMPBELL
University of California, Berkeley

SITUATIONS frequently occur in which there are available a fairly large number of variables, such as a set of criterion measures, and the researcher wishes to identify the dimensions of common variance. Various methods of cluster or factor analysis (Harman, 1960; Tryon, 1964) are available for handling this situation but they all require that each individual be measured on an appreciable number of the variables. Unfortunately, this situation does not always exist in the real world and a complete covariation matrix cannot be obtained. An extreme example of such a situation would be when each individual is measured on only one variable or criterion. Under these conditions no indices of covariation could be obtained for pairs of variables and a cluster or factor analysis could not be performed. In hopes of offering an answer to the above dilemma the present investigation uses and evaluates a technique for clustering variables based on the similarity of their regression equations rather than their intercorrelations. Since it is not dependent on the intercorrelations between criteria, it thus might offer some hope for overcoming an all too prevalent constraint.

Method

Bottenberg and Cristal (1961) have presented a technique for hierarchically clustering a set of regression equations into optimally predictable groups such that a smaller number of regression

equations can be used to predict the individual criteria with minimal loss of predictive efficiency. A computer program for carrying out this procedure has been developed by Ward (1961). From examination of expressions (1) and (2) below it can be seen that this technique accomplishes its purpose by grouping together homogeneous or similar regression equations. For purposes of this investigation it seemed reasonable to assume that if two criteria are highly intercorrelated then their individual regression equations, when a common set of predictors is used and when the multiple R is reasonably high, should tend to "look alike." That is, the individual predictors should have similar correlations with the two criteria and their standard partial regression weights should be similar for the two equations. Criteria could then be clustered on the basis of the relative similarity of their regression equations.

The Bottenberg and Cristal procedure begins by computing a weighted average of the individual multiple correlation coefficients using the following expression.

$$o^{R^2}(g) = \frac{\dfrac{\sum_{i=1}^{g} n_i R_i^2 s_i^2}{N} + \dfrac{\sum_{i=1}^{g} n_i \bar{X}_i^2}{N} - \bar{X}_.^2}{S^2} \tag{1}$$

Where: g = the total number of criterion groups (including ungrouped individual criteria).

n_i = number of observations in criterion group i.

R_i^2 = squared multiple correlation of predictor battery with criterion group i.

s_i^2 = variance of criterion scores in group i.

$\bar{X}_i^2$ = mean of criterion scores in group i.

N = total number of individuals in all criterion groups.

$\bar{X}_.$ = grand criterion mean.

S^2 = grand criterion variance.

The two individual equations are then combined which reduce this over-all average R^2 the *least*. To do this, every possible pair of the original set of g equations is evaluated by means of the following expression.

$$ {}_cR_{(m)}^2 = \frac{\frac{1}{N^2} \sum_{i=1}^{m} \sum_{h=1}^{m} n_i n_h s_i^2 s_h^2 \cdot \left(\sum_{j=1}^{p} B_{ij} v_{hj} \right)}{S^2} \tag{2} $$

Where: m = the number of criteria in the cluster.

B_{ij} = standardized regression weight for the jth predictor of criterion i.

v_{hj} = corresponding validity coefficient for the jth predictor of criterion h.

s_i^2 and s_h^2 = variance of the observations in the individual criterion groups.

n_i and n_h = number of observations in the individual criterion groups.

N = sum of the observations in the I criterion groups.

The subscripts i and h both take on all possible values. In the case of evaluating pairs of equations the double summation in expression (2) will be over four different values. ${}_cR_{(m)}^2$ is thus an average quantity obtained by using all possible combinations of zero-order correlations and standardized partial regression coefficients. The more similar the correlations and Beta weights for two criteria, the smaller will be the relative decrement in ${}_cR_{(m)}^2$. In other words, by means of expression (2) the total number of multiple correlation coefficients is reduced from g to g-1 and that combination is made which minimizes the difference between the average of g ${}_cR_{(g)}^2$'s and $g - 1$ ${}_cR_{(g)}$'s. The highest average will always be obtained when an individual equation is used for each criterion. Criteria are clustered one at a time, and this hierarchical grouping procedure is continued until all variables or criteria have been combined into one group and ${}_cR_{(g)}^2$ is at a minimum.

In order to make a one-to-one comparison between the multiple regression method and an analysis which utilizes a covariation matrix, the Holzinger-Harman (Fruchter, 1954) method of cluster analysis was modified so that variables were clustered in a hierarchical fashion and the variables within clusters were as homogeneous as possible with regard to their intercorrelations. With this method the index of cluster homogeneity is the B-coefficient, or ratio of the average of the intercorrelations of the variables within a

cluster to the average of the correlations of the cluster variables with the variables outside the cluster. Variables were clustered one at a time by computing the B-coefficient for every possible combination and making that combination which maximized the B-coefficient or average of the B-coefficients at each stage. This procedure starts by clustering the two variables which have the highest intercorrelation. The second step involves either clustering a third variable with the initial pair or clustering two variables to form a second pair. The decision is dependent on which is larger, the B-coefficient for the three variable cluster or the average of the B-coefficients for the two pairs. Variables are clustered in this fashion, one at a time, until all variables are grouped into clusters. These clusters are then grouped to form larger clusters until all variables have been clustered into one group. Thus the number of stages corresponds to the number of variables, and at each stage the most homogeneous grouping is achieved in terms of the variable intercorrelations.

Subjects

The individuals used in the study were 406 students in the University of Minnesota's Institute of Technology who had completed the first two years of the engineering curriculum.

Predictors and Criteria

Five predictors were included: (1) high school rank, (2) the Minnesota Scholastic Aptitude Test (MSAT), (3) the Mechanics of Expression score and (4) the Effectiveness of Expression score from the *Cooperative English Test: Lower and Higher Levels*, and (5) the Institute of Technology Mathematics Test (IT Math). Predictors two, three, and four are regularly given to Minnesota high school juniors as part of the university counseling bureau's statewide testing program (Berdie, et al, 1959). The IT Math test is given to all high school seniors who apply for admittance to the Institute of Technology.

The criteria consisted of twenty of the basic core courses taken by almost every engineering student during his freshman and sophomore years. They are listed in Table 1.

TABLE 1

The Basic Set of Twenty Core Courses in the Engineering Curriculum That Were Used as Criteria

1) Composition 1A	(Comp 1)	10) Calculus I	(Calc 1)
2) Composition 2A	(Comp 2)	11) Calculus II	(Calc 2)
3) Composition 3A	(Comp 3)	12) Calculus III	(Calc 3)
4) Literature 1A	(Lit 1)	13) Calculus IV	(Calc 4)
5) Literature 2A	(Lit 2)	14) General physics 11	(Phys 11)
6) Literature 3A	(Lit 3)	15) General physics 12	(Phys 12)
7) Engineering graphics 14	(EG 14)	16) General physics 13	(Phys 13)
8) Engineering graphics 15	(EG 15)	17) General physics 14	(Phys 14)
9) Algebra, trigonometry, and analytic geometry	(Algebra)	18) Intermediate physics 50	(Phys 50)
		19) General chemistry 15	(Chem 15)
		20) General chemistry 16	(Chem 16)

Procedure

To check the relative stability of the two solutions the sample was split into two groups (A and B) of 203 individuals each and the analysis carried out in identical fashion on both groups. For each subsample the regression equations for each of the criteria and the criterion intercorrelation matrix were computed. Both the multiple regression and the cluster analysis technique were then used to cluster hierarchially the individual criteria. At each stage, for both methods, the multiple correlation of the predictor battery with each total cluster score and the intercorrelations of the cluster scores were obtained.

Results

The criterion clusters obtained by each method at selected stages are shown in Tables 2 and 3 for subsamples A and B respectively. As can be seen from these two tables the two methods tend to give dissimilar results when applied to the same data. This lack of similarity can best be illustrated by an attempt to conceptualize and name the clusters at say stages 15 through 19.

For the analysis of intercorrelations the task is not too difficult. All the English courses group together in an apparent verbal factor which includes both the ability to write clearly and correctly (comprehension) and to understand what has been read (literature). By stage 15 this cluster is complete and it does not change in successive stages. The two engineering graphics courses also group

TABLE 2

Comparative Results of Hierarchical Clustering of Criteria on the Basis of Both the Homogeneity of Regression Equations and the Criterion Intercorrelations

Sample A						
Stage	By regression equation			By intercorrelation		
(1)	All criteria treated individually					
(2)	Comp 1 Comp 2 (.43)			Comp 3 Lit 3 (.48)		
(3)	Comp 1 Comp 2 (.43)	Phys 12 Phys 13 (.41)		Comp 3 Lit 3 Comp 2 (.49)		
(4)	Comp 1 Comp 2 (.43)	Phys 12 Phys 13 (.41)	Comp 3 Lit 3 (.48)	Comp 3 Lit 3 Comp 2 Lit 2 (.49)		
(5)	Comp 1 Comp 2 (.43)	Phys 12 Phys 13 (.41)	Comp 3 Lit 3 (.48)	Comp 3 Lit 3 Comp 2 Lit 2 Comp 1 (.50)		
	EG 14 Chem 15 (.45)					
(10)	Comp 1 Comp 2 Lit 2 (.45)	Phys 12 Phys 13 (.41)	Comp 3 Lit 3 (.48)	Comp 3 Lit 3 Comp 2 Lit 2 Comp 1 Lit 1 (.51)	Chem 15 Chem 16 (.38)	Phys 12 Phys 13 Phys 11 Algebra (.52)
	EG 14 Chem 15 Phys 11 (.55)	Calc 3 Phys 50 (.29)	Calc 2 Phys 14 (.33)			
	Calc 1 Chem 16 (.42)					
(15)	Comp 1 Comp 2 Lit 2 Lit 1 Comp 3 Lit 3 (.51)	Phys 12 Phys 13 EG 14 Chem 15 Phys 11 (.53)		Comp 3 Lit 3 Comp 2 Lit 2 Comp 1 Lit 1 (.51)	Chem 15 Chem 16 (.38)	Phys 12 Phys 13 Phys 11 Algebra Phys 14 (.50)
	Calc 3 Phys 50 Calc 1 Chem 16 (.39)	Calc 2 Phys 14 Algebra (.44)		EG 14 EG 15 (.37) Diff-E	Calc 2 Calc 3 Calc 1 (.39)	
(16)	Comp 1 Comp 2 Lit 2 Lit 1	Phys 12 Phys 13 EG 14 Chem 15	Calc 2 Phys 14 Algebra Calc 3	Comp 3 Lit 3 Comp 2 Lit 2	Chem 15 Chem 16 (.38)	Phys 12 Phys 13 Phys 11 Algebra

TABLE 2—*Continued*

Sample A						
Stage	By regression equation			By intercorrelation		
	Comp 3 Lit 3 (.51)	Phys 11 (.53)	Phys 50 Calc 1 Chem 16 (.43)	Comp 1 Lit 1 (.51) EG 14 EG 15 (.37)	Calc 2 Calc 3 Calc 1 Diff-E Phys 50 (.37)	Phys 14 (.50)
(17)	Comp 1 Comp 2 Lit 2 Lit 1 Comp 3 Lit 3 (.51)	Phys 12 Phys 13 EG 14 Chem 15 Phys 11 EG 15 (.51)	Calc 2 Phys 14 Algebra Calc 3 Phys 50 Calc 1 Chem 16 (.43)	Comp 3 Lit 3 Comp 2 Lit 2 Comp 1 Lit 1 (.51) EG 14 EG 15 (.37)	Chem 15 Chem 16 (.38)	Phys 12 Phys 13 Phys 11 Algebra Phys 14 Calc 2 Calc 3 Calc 1 Diff-E Phys 50 (.47)
(18)	Comp 1 Comp 2 Lit 2 Lit 1 Comp 3 Lit 3 Diff-E (.47)	Phys 12 Phys 13 EG 14 Chem 15 Phys 11 EG 15 (.51)	Calc 2 Phys 14 Algebra Calc 3 Phys 50 Calc 1 Chem 16 (.43)	Comp 3 Lit 3 Comp 2 Lit 2 Comp 1 Lit 1 (.51) EG 14 EG 15 (.37)	Chem 15 Chem 16 Phys 12 Phys 13 Phys 11 Algebra Phys 14 Calc 2 Calc 3 Calc 1 Diff-E Phys 50 (.48)	
(19)	All English courses + Diff-E (.47)	All non-English courses except Diff-E (.49)		All English courses (.51)	All non-English courses (.49)	
(20)	All courses combined (.52)			All courses combined (.52)		

Note—Only those criteria which were clustered at Stages 1–5, 10, and 15–20 are shown. The multiple correlations of the predictor battery with the criterion cluster score is given in parentheses.

together. At stage 15 all the calculus courses and the courses in differential equations group together, but the course in algebra clusters with the physics courses. On the basis of course content the calculus courses and differential equations would seem to represent a higher mathematics factor that requires a much more abstract

TABLE 3

Comparative Results of Hierarchical Clustering of Criteria on the Basis of Both the Homogeneity of Regression Equations and the Criterion Intercorrelations

Sample B						
Stage	By regression equation			By intercorrelation		
(1)	All criteria treated individually					
(2)	Comp 2 Lit 1 (.43)			Comp 2 Lit 2 (.43)		
(3)	Comp 2 Lit 1 (.43)	Calc 3 Phys 50 (.40)		Comp 2 Lit 2 (.43)	EG 14 EG 15 (.36)	
(4)	Comp 2 Lit 1 (.43)	Calc 3 Phys 50 (.40)	Comp 3 Lit 2 (.48)	Comp 2 Lit 2 Comp 1 (.48)	EG 14 EG 15 (.36)	
(5)	Comp 2 Lit 1 (.43)	Calc 3 Phys 50 (.40)	Comp 3 Lit 2 (.48)	Comp 2 Lit 2 Comp 1 Lit 1 (.48)	EG 14 EG 15 (.36)	
	Calc 2 Phys 14 (.46)					
(10)	Comp 2 Lit 1 Comp 1 (.47)	Calc 3 Phys 50 Calc 2 Phys 14 (.42)	Comp 3 Lit 2 (.48)	Comp 2 Lit 2 Comp Lit 1 Comp 3 Lit 3 (.53)	EG 14 EG 15 (.36)	Phys 12 Phys 13 Phys 11 (.51)
	EG 14 Phys 12 (.46)	Algebra Calc 1 (.53)	Phys 11 Chem 15 (.54)		Calc 3 Diff-E (.37)	
(15)	Comp 2 Lit 1 Comp 1 Lit 3 Comp 3 Lit 2 (.53)	Calc 3 Phys 50 Calc 2 Phys 14 Diff-E (.43)	EG 14 Phys 12 Phys 11 Chem 15 (.54)	Comp 2 Lit 2 Comp 1 Lit 1 Comp 3 Lit 3 (.53)	EG 14 EG 15 (.36)	Phys 12 Phys 13 Phys 11 Algebra (.54)
	Algebra Calc 1 Chem 16 (.57)			Calc 3 Diff-E Calc 1 Calc 2 (.46)	Chem 15 Chem 16 (.51)	Phys 14 Phys 50 (.39)
(16)	Comp 2 Lit 2 Comp 1 Lit 3 Comp 3 Lit 2 (.53)	Calc 3 Phys 50 Calc 2 Phys 14 Diff-E (.43)	EG 14 Phys 12 Phys 11 Chem 15 (.54)	Comp 2 Lit 2 Comp 1 Lit 1 Comp 3 Lit 3 (.53)	EG 14 EG 15 (.36)	Phys 12 Phys 13 Phys 11 Algebra (.54)
	Algebra Calc 1	EG 15 Phys 13		Calc 3 Diff-E	Chem 15 Chem 16	

TABLE 3—*Continued*

	Sample B					
Stage	By regression equation			By intercorrelation		
	Chem 16 (.57)	(.44)		Calc 1 Calc 2 (.46)	Phys 14 Phys 50 (.49)	
(17)	Comp 2 Lit 2 Comp 1 Lit 3 Comp 3 Lit 2 (.53)	Calc 3 Phys 50 Calc 2 Phys 14 Diff-E (.43)	EG 14 Phys 12 Phys 11 Chem 15 Algebra Calc 1 Chem 16 (.59)	Comp 2 Lit 2 Comp 1 Lit 1 Comp 3 Lit 3 (.53)	EG 14 EG 15 (.36)	Phys 12 Phys 13 Phys 11 Algebra Calc 3 Diff-E Calc 1 Calc 2 (.53)
	EG 15 Phys 13 (.44)				Chem 15 Chem 16 Phys 14 Phys 50 (.49)	
(18)	Comp 2 Lit 2 Comp 1 Lit 3 Comp 3 Lit 2 (.53)	Calc 3 Phys 50 Calc 2 Phys 14 Diff-E (.43)	EG 14 Phys 12 Phys 11 Chem 15 Algebra Calc 1 Chem 16 EG 15 Phys 13 (.57)	Comp 2 Lit 2 Comp 1 Lit 1 Comp 3 Lit 3 (.53)	EG 14 EG 15 (.36)	Phys 12 Phys 13 Phys 11 Algebra Calc 3 Diff-E Calc 1 Calc 2 Chem 15 Chem 16 Phys 14 Phys 50 (.53)
(19)	All English courses (.53)	All non-English courses (.54)		All English courses (.53)	All non-English courses (.54)	
(20)	All courses combined (.59)			All courses combined (.59)		

Note—Only those criteria which were clustered at Stages 1–5, 10, and 15–20 are shown. The multiple correlations of the predictor battery with the criterion cluster score is given in parentheses.

and qualitatively different kind of problem solving from that in the course in algebra. It is not too surprising that this course groups with the first three physics courses, since many of the skills required to solve introductory physics problems are algebraic in nature. The two chemistry courses, which also group together at stage 15, remain as a discrete cluster until they group with the general mathematic's-science factor at a later stage. The only apparent ambiguity in these results is the behavior of the two more advanced physics courses. In subsample B they form a separate cluster and then group with the chemistry courses. However, in subsample A, phys-

ics 14 clusters with algebra and introductory physics while physics 50 clusters with the higher mathematics courses. Again however, the content of these last two physics courses is qualitatively different (more advanced topics) from the more basic introductory courses taken during the first year of the engineering curriculum.

The behavior of these clusters across successive stages also makes reasonable sense. The higher mathematic's factor merges with the physics factor and then with the chemistry factor to form a general mathematic's-science factor. The spatial relations factor remains distinct until at stage 19 it is forced to choose between the verbal factor and the mathematic's-science factor and as would be expected, merges with the latter. Again, the only exception to this conceptualization is the clustering behavior of physics 14 and physics 50.

It is much more difficult to attach a name and conceptual description to the clusters obtained by grouping criteria with homogeneous regression equations. For example, at stage 15 for subsample A, the chemistry courses are split between two clusters and so are the mathematical courses. Although the first cluster contains all the English courses, the second cluster includes the first course in chemistry and engineering graphics and the first three courses in physics; the third cluster contains algebra, physics 14, and calculus 2; and the fourth cluster is made up of the last course in physics, chemistry, and calculus, plus the beginning course in calculus. It is almost impossible to distinguish among these last three clusters in terms of criterion content. The situation does not improve much across successive stages. It is made even more difficult when the course in differential equations clusters with the English courses at stage 18!

In subsample B, at stages 15 and 16, the situation is not any clearer than in subsample A. At stage 18 the picture is somewhat better than it was in subsample A, since differential equations does not cluster with the English courses and the advanced mathematic and physics group together in what might possibly be called an abstract quantitative reasoning factor. However, the third cluster remains very difficult to name or define.

Examination of Tables 2 and 3 also indicates differing degrees of stability across subsamples for the clusters obtained by the two methods. For the intercorrelation method the only discrepancy

across subsamples is the clustering of physics 14 and 50 with the chemistry courses in subsample B, but with two different clusters in subsample A. By contrast, there is a much greater number of discrepancies between subsamples for the regression analysis.

As might be expected the average intercorrelation of the total cluster scores obtained by grouping regression equations tends to be somewhat higher than that for the intercorrelation method. Table 4 shows this comparison for stage 18. At this stage in the clustering process all 20 of the criteria have been grouped by each method, and the number of criterion groups are the same.

TABLE 4

Intercorrelations of Cluster Scores at Stage 18 for Both Methods and for Both Subsamples

Group	Cluster Score Intercorrelations					
	Regression Analysis			Correlation Analysis		
	1	*2*	*3*	*1*	*2*	*3*
A		.49	.63		.50	.34
			.73			.48
	$\bar{r}$ = .63[a]			$\bar{r}$ = .44		
B		.36	.41		.21	.42
			.78			.48
	$\bar{r}$ = .55			$\bar{r}$ = .37		

[a] The average of each set of cluster score intercorrelations was computed by means of Fisher's Z transformation.

Discussion and Conclusions

The two methods apparently do *not* give similar results, at least within the limitations of the present study. Although there is no objective decision rule that can be used to support this contention, it seems obvious that the clusters obtained at stages 15, 16, 17, for example, would be conceptualized differently depending on which method was used. Furthermore, the clusters obtained from the analysis of intercorrelations make considerably more "sense" in terms of apparent criterion content. The results from this method also appear to be more stable across the two subsamples.

The above situation is somewhat analogous to the problem of estimating the factor loadings of a set of "outside" variables from the factor matrix obtained on another set of variables. Techniques for doing this (Dwyer, 1937) also required the initial correlation of each outside variable with every variable in the matrix (in this

study, the correlation of each criterion with every predictor). However, this type of procedure is most applicable to the situation where the original set of variables is large and the number of outside variables is small. The more usual situation in applied prediction is the reverse, or the situation embodied in this paper. The limiting factor in both situations is the degree to which the original set of variables (predictors in this case) account for the common variance in the outside variables. Apparently predictors which yield multiple correlations of this magnitude, which seem to be typical for the usual real world situation, do not account for enough of the common variance to make a clustering of regression equations trustworthy as a substitute for an unobtainable covariation matrix.

Independent of whether or not an index of regression equation homogeneity can be substituted in a relative manner for an intercorrelation, is the question of which set of clusters in the hierarchical solution to use for purposes of obtaining cluster scores or making statements about criterion dimension. Tryon (1964) makes the point that the number of clusters or factors extracted from a set of "real" intercorrelations is somewhat arbitrary. The number cannot be decided on the basis of the "rank" of the matrix, since in order for all the residuals to become zero the number of factors must equal the number of variables. Factoring or clustering is merely continued until a relevant index such as the average squared residual or average obtained communality reaches a certain value.

If the aim of the researcher is to obtain clusters of variables which define a reduced number of dimensions and yield useful cluster scores, it would seem that obtaining maximally homogeneous subsets in a hierarchical fashion would provide a considerable amount of information. In the present example it would certainly be possible to compute some index of factoring such as the average squared residual at each stage. It is also possible to examine the relative homogeneity of the clusters, the intercorrelations among cluster scores, and the predictability (multiple-R with predictor battery) of each cluster score. By utilizing the hierarchical procedure the changing pattern of all these indices can be noted. For many research purposes the reduction of the residuals to some specified level may not be the most important consideration, and a solution which optimized the independence and predictability of

simple sum cluster scores would be preferred. Thus for many applied problems a procedure which gives a picture of the changing pattern of these indices would seem to be very valuable. The investigator could then choose the pattern most advantageous to him.

Summary

The present study compared two methods of grouping or clustering criteria into homogeneous subsets in hopes of finding a solution to the problem of incomplete covariation matrices. A technique developed by Bottenberg and Cristal was used to group criteria on the basis of the homogeneity or similarity of their regression equations, given a common set of predictors. This technique was compared to the Holzinger-Harman method of cluster analyzing the intercorrelation matrix. The multiple regression technique does not depend on the intercorrelations of the criteria. Both methods proceed by clustering one variable at a time in a stepwise fashion. The solutions obtained by the two methods were judged not to be comparable; however, certain advantages of the hierarchical method of clustering were pointed out.

REFERENCES

Berdie, R. F., Layton, W. L., Swanson, E. O., and Hagenah, Theda. *Counseling and the Use of Tests: A Manual for the Minnesota State-Wide Testing Program*. Minneapolis: University of Minnesota Counseling Bureau, 1959.

Bottenberg, R. A. and Cristal, R. E. *An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency*. Lackland Air Force Base: Personnel Laboratory, Wright Air Development Division—Air Research and Development Command (Technical Note WADD-TN-61-30), March 1961.

Dwyer, P. S. The Determination of Factor Loadings of a Given Test from the Known Factor Loadings of Other Tests. *Psychometrika*, 1937, 2, 173–178.

Fruchter, B. *Introduction to Factor Analysis*. New York: D. Van Nostrand, 1954.

Harman, H. H. *Modern Factor Analysis*. Chicago: The University of Chicago Press, 1960.

Tryon, R. C. *The Component Programs of the BC TRY System*. (Mimeographed report) Berkeley: University of California, 1964.

Ward, J. H., Jr. *Hierarchical Grouping to Maximize Payoff*. Lackland Air Force Base: Personnel Laboratory, Wright Air Development Division—Air Research and Development Command (Technical Note WADD-TN-61-29), March 1961.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

AN INTERPRETATION OF THE COEFFICIENTS OF PREDICTIVE VALIDITY AND OF DETERMINATION IN TERMS OF THE PROPORTIONS OF CORRECT INCLUSIONS OR EXCLUSIONS IN CELLS OF A FOURFOLD TABLE

WILLIAM B. MICHAEL
University of California, Santa Barbara

NUMEROUS and varied approaches have been presented in statistics textbooks and journal articles to explain what is meant by the correlation coefficient or the coefficient of determination. Among the most helpful expositions for the understanding of the meaning of predictive validity coefficients was that of Taylor and Russell (1939) who not only proposed the selection ratio but also furnished useful charts to indicate the proportions of examinees who would be successful, relative to (a) several cutting scores in the predictor variable, (b) different proportions of employees considered successful on the job, and (c) various values for a validity coefficient. Using the normal bivariate surface as a model, Guilford and Michael (1949) extended Taylor and Russell's work and furnished charts that showed the relationship between the proportion of those selected who would be successful and the selection ratio (proportion selected) for specified proportions of individuals successful on the criterion and for different validity coefficients. A chapter presented by Ghiselli and Brown (1955, pp. 140–48) and another chapter prepared by Thorndike and Hagen (1960, pp. 169–72) have also furnished useful indices and charts that provide helpful insights in the interpretation of validity coefficients. Thorndike and Hagen's presentation probably comes closest to the approach that will be developed in this paper.

Problem

Despite the aid furnished by these sources, beginning students in courses in psychological statistics and measurement encounter considerable difficulty in understanding either the predictive validity coefficient or its squared value, the coefficient of determination. It is the purpose of the writer to present three simple indices as well as a chart to accompany one of them which in his teaching experience have appeared to aid students in grasping what is meant by a predictive validity coefficient when the normal correlation surface is subdivided at the median in both the predictor and criterion variable.

The underlying rationale involves essentially a comparison of the proportion of individuals who are correctly placed on the criterion by the predictor with (a) the size of the validity coefficient r and (b) the value of the corresponding coefficient of determination r^2. A corollary to this rationale is a comparison of the proportion of individuals correctly placed in excess of a chance proportion of .50 (when $r = 0$) with the magnitude of both r and r^2.

Definitions of Terms

In Figure 1 the fourfold table sets forth the definitions of the principal terms involved in the median dichotomization of the normal correlation surface on the predictor and criterion dimensions. The symbols A, B, D and C stand for the numbers of individuals in the first, second, third, and fourth quadrants, respectively. The individuals in the first and third quadrants represented by the sum $A + D$ constitute those who have been correctly placed or who may be designated as "hits." Thus the frequency A stands for examinees who are above the median in both the predictor and criterion variables, and D for those below the median on each of these two variables. The frequencies B and C in the second and fourth quadrants, respectively, designate individuals who have been incorrectly placed—the "misses," since their relative position on the criterion variable is contrary to that on the predictor variable. In the normal bivariate model the sum $A + D$ should be equal to $B + C$ when the correlation is zero. As the correlation increases in positive value for a population of fixed size N, the frequency $A + D$ becomes monotonically higher than the frequency $B + C$,

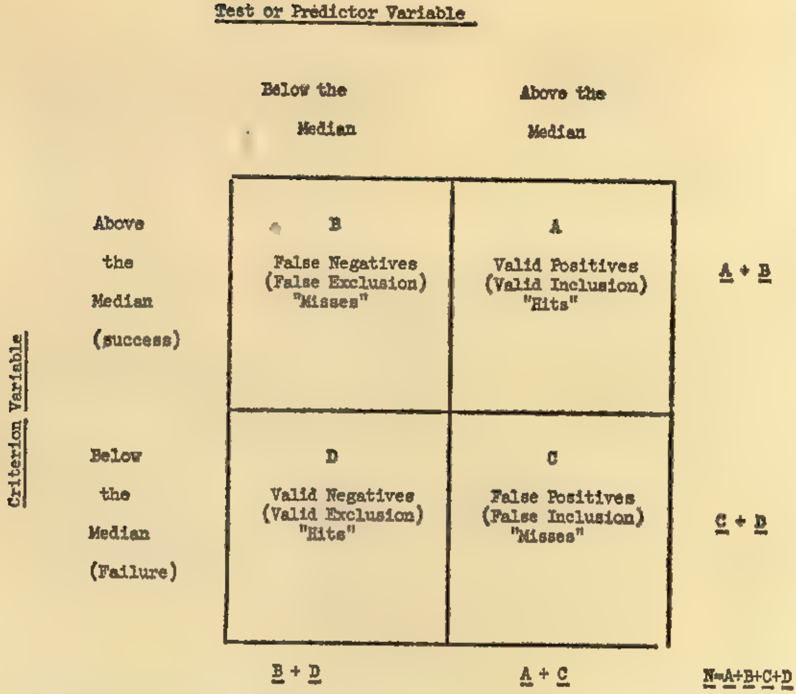


Figure 1. Correlation surface subdivided by a medium test score which separates the population into low- and high-scoring groups of individuals and by a median criterion value which divides the same population into their successful and unsuccessful standing in job performance.

which is becoming progressively smaller. (When the correlation is negative, the sum of $B + C$ will exceed that of $A + D$).

Each of the three terms which serve to designate the entries in the third, fourth, and fifth columns of Table 1 may be defined for $r \geqq 0$ as follows:

$H = \dfrac{A + D}{N} =$ the proportion of individuals in the total population of size N who are correctly placed—the proportion of "hits";

$E = \dfrac{A + D}{N} - .5000 =$ the proportion of correctly placed individuals in excess of the chance proportion of .50 which is associated with an r of zero; and

$$I = \frac{100\left(\frac{A+D}{N}\right) - .5000}{.5000}$$ = the percentage of improvement in correct (valid) placements relative to the proportion of .50 for an r of zero.

Similar definitions could be formed with $B + C$ taking the place of $A + D$ when r is negative. Although not presented, ratios of $A + D$ to $B + C$ or of $B + C$ to $A + D$ also furnish bases for interpretation of a correlation coefficient.

TABLE 1

Three Indices Furnishing Additional Information Regarding the Meaning of Predictive Validity for Various Values in r and r^2 When a Correlation Surface is Subdivided at the Median in Both the Predictor and Criterion Variables

Correlation Coefficient r	Coefficient of Determination r^2	Proportion of Correct (Valid) Placements (H) $\frac{A+D}{N}$	Proportion Correctly Placed in Excess of Chance Proportion of .50 for $r = 0.00$ (E) $\frac{A+D}{N} - .5000$	Percentage of Improvement in Correct (Valid) Placements Relative to Proportion of .50 for $r = 0.00$ (I) $\frac{100\left(\frac{A+D}{N}\right) - .5000}{.5000}$
.00	.0000	.5000	.0000	0.00
.05	.0025	.5159	.0159	3.18
.10	.0100	.5319	.0319	6.38
.15	.0225	.5479	.0479	9.59
.20	.0400	.5641	.0641	12.82
.25	.0625	.5804	.0804	16.09
.30	.0900	.5970	.0970	19.40
.35	.1225	.6138	.1138	22.76
.40	.1600	.6310	.1310	26.20
.45	.2025	.6486	.1486	29.72
.50	.2500	.6667	.1667	33.33
.55	.3025	.6854	.1854	37.07
.60	.3600	.7048	.2048	40.97
.65	.4225	.7252	.2252	45.05
.70	.4900	.7468	.2468	49.36
.75	.5625	.7699	.2699	53.99
.80	.6400	.7952	.2952	59.03
.85	.7225	.8234	.3234	64.68
.90	.8100	.8564	.3464	71.29
.95	.9025	.8989	.3989	79.78
1.00	1.0000	1.0000	.5000	100.00

Procedure

The H entries in Table 1 were obtained from Pearson's (1931) "Volumes of a Bivariate Surface," and the E and I values were quickly derived from the H indices. As mentioned previously, the normal bivariate model was assumed to be appropriate.

Findings

From the entries in Table 1, one can quickly ascertain the nature of the relationship of the indices H, E, and I to both r and r^2. For example, when $r = .60$, or $r^2 = .36$, it is noted that the proportion H of valid positives and valid negatives—the proportion of correctly placed individuals—is .7048. Thus on the criterion, the standing of slightly more than 7 out of 10 individuals is being accurately predicted. The complementary proportion $1 - H$ of .2952 constitutes the proportion of "misses", or incorrect placements. The index E with its value of .2048 for an r of .60 simply reflects the proportion in the total population accurately placed in excess of the chance proportion of .5000 which would be expected when the correlation is zero. This proportion of .2048 may also be viewed as representing an improvement I of 40.97 per cent (with a selection ratio of 1.00) in correct or valid placements over what would be the number of correct placements when the correlation is zero.

It is also interesting to note that by interpolation a coefficient of about .71, which is associated with an r^2 of nearly .50, also corresponds to a coefficient H of about .75—a value which represents an absolute gain of about .25 in the proportion E of valid inclusions and valid exclusions—or a magnitude in the improvement index I of nearly 50 per cent in terms of the number of accurate placements of individuals over the number when the correlation is zero. It can be seen that in the instance of a value of .71 in r, and in this instance alone, r^2 is approximately .50, which is equal to .01 I. In other words, for an r of about .71, the coefficient of determination converted to a percentage corresponds to the percentage of improvement I (approximately 50) in valid placements which takes place between an r of 0.00 and an r of 1.00.

Since a coefficient of .71 is close to the upper limit of predictive validities in most personnel selection programs, it is apparent that for the *median splits* involved the proportion of correct placements

is not likely to exceed .75. Thus without a testing device, one could choose 2 out of 4 individuals correctly, but under optimal conditions in practical selection with tests, the greatest accuracy to be anticipated is that one could select 3 out of 4 individuals correctly when the selection ratio is 1.00 and when the proportion of individuals successful on the criterion is .50. (Of course, with use of small selection ratios such as .10 or .20, proportions of correct placements could often approximate .80 or .85 for highly valid tests and for relatively difficult jobs as defined by rather small percentages of 15 to 25 of individuals who have been judged to be successful on the criterion measure.)

In Figure 2 the relationship between H and r as well as between H and r^2 is portrayed. Thus it is apparent that for the customary range of predictive validity coefficients between .30 and .65 the proportion of correct placements H varies between about .60 and .73. The complementary proportion of incorrect placements, or "misses," known as false negatives and false positives, is given by 1 — .60 and 1 — .73 respectively, or by .40 and .27. The value of H is represented by the ordinate from the base line, whereas 1-H is given by the vertical distance from the horizontal boundary line at the top of Figure 2.

Summary

Largely for pedagogical purposes three indices have been proposed to explain the meaning of predictive validity. Use was made of the normal bivariate surface which was dichotomized at the median in both the predictor and criterion variables. The basic approach was that of relating three indices derived from the combined proportions of individuals in two cells along a diagonal of the fourfold table to values of the predictive validity coefficient r and of its coefficient of determination r^2.

REFERENCES

Ghiselli, Edwin E. and Brown, C. W. *Personnel and Industrial Psychology*. (2ed.) New York: McGraw Hill Book Co., 1955.

Guilford, J. P. and Michael, William B. *The Prediction of Categories from Continuous Measurements: Applications to Personnel Selection and Clinical Prognosis*. Beverly Hills, California: Sheridan Supply Co., 1949.

Pearson, Karl. *Tables for Statisticians and Biometricians*, Part II. England: Cambridge University Press, 1931.

Taylor, H. C. and Russell, J. T. The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables. *Journal of Applied Psychology,* 1939, 23, 565–578.

Thorndike, Robert L. and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education.* (2ed.) New York: John Wiley & Sons, Inc., 1961.

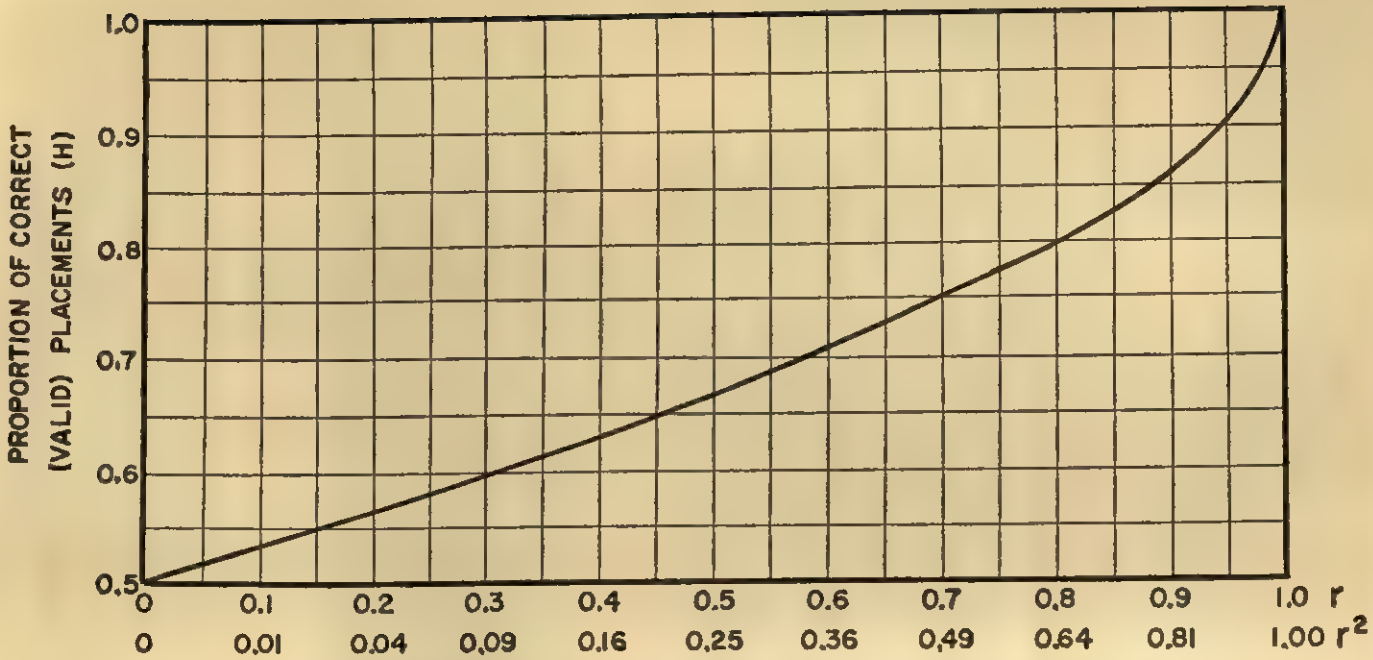


Figure 2. Proportion of correct (valid) placements (H) in a fourfold table with median splits in the marginal variables as a function of the size of the predictive validity coefficient r and of its coefficient of determination r^2.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

THE PREDICTIVE RELATIONSHIP OF THE MILLER ANALOGIES TEST TO OBJECTIVE AND SUBJECTIVE CRITERIA OF SUCCESS IN A GRADUATE SCHOOL OF EDUCATION[1]

DAVID A. PAYNE and CYNTHIA E. TUTTLE
Syracuse University

The Miller Analogies Test (MAT) continues to be one of the most widely used instruments for graduate school selection. A review of MAT research reveals a wide range of validity coefficients. Values from .00 (Travers, 1948) to .69 (Gustad, 1950) are noted. Such discrepancies can in part be attributed to the considerable variation in candidate qualifications and institutional differences. The lack of reliable success criteria is also a significant contributor. The usual criterion of grades is generally conceded to lack sufficient consistency to allow for stable prediction. Very few validity studies can be located which involve the use of objective criteria, such as standardized instruments or end-of-program comprehensive examinations. The purpose of this paper is to report data which compare the predictability of the MAT relative to the criteria of grades and comprehensive examination scores.

Variables

The course work leading to a Masters Degree in Education at a large urban university requires, in addition to the usual electives and prescribed major instruction, a minimum of one course in each of the following areas, Educational Psychology (EP), Meas-

[1] Data analysis was completed with support of National Science Foundation Grant GP-1137 made to Syracuse University Computing Center. The assistance of Miss Agnes R. Esposito and Mr. James G. Black in the completion of this study is gratefully acknowledged.

urement and Statistics (MS), and Cultural Foundations of Education (CF). Students frequently take more than one course in each of the core areas. The following subjective estimates of success (using a threepoint scale) were derived from each student's transcript:

1. Mean grade in Educational Psychology core courses (EPG)
2. Mean grade in Measurement and Statistics core courses (MSG)
3. Mean grade in Cultural Foundations (CPG)
4. Total Master's grade point average (TG)

Chansky (1964) recently noted that since grades do not meet the assumptions underlying a ratio or interval scale of measurement, the "mean GPA" tends to be relatively meaningless. He suggested that a median GPA be used whenever grades were used as criteria. The final grade criterion was, therefore:

5. Median grade in total Master's program (MdTG)

Master's candidates are required to pass a 285-item comprehensive examination which equally represent the three core areas. Approximately 80 per cent of the items are multiple-choice; the remainder, true-false and matching. The examination was judged to have high content validity, as the items were written by the core area instructors and then evaluated by test construction experts. Part and total scores formed the basis for four more criteria (with Kuder Richardson Formula 21 reliabilities—subsequently abbreviated as KR_{21}—in parentheses):

6. Educational Psychology subtest score	(EPS)	(.47)
7. Measurement and Statistics subtest score	(MSS)	(.82)
8. Cultural Foundations subtest score	(CFS)	(.65)
9. Total Master's Comprehensive Examination Score	(TMCS)	(.84)

Form K of the Miller Analogies Test (MAT) ($KR_{21} = .91$) was administered when students applied for graduate work. The distribution of times between MAT and completion of degree was positively skewed, with a mean of 23 months and a median of 20 months.

Sample

The sample was composed of 115 males and 104 females. Since no sex differences were noted, the groups were combined ($N = 219$). Mean age of the total group was 30 years. All students not only had taken the MAT but also had completed the Master's degree between July 1958 and March 1963.

Results and Discussion

Means, standard deviations, and correlations of variables involved in the study are presented in Table 1. The greater effectiveness of the MAT in predicting Comprehensive Examination scores than grades is noted. This finding, in part, could be expected when one considers the skills required for successful performance in the classroom versus those involved in objective written examinations. The difference in validity coefficients reflects, perhaps, the lack of uniform grading practices. Such a lack is particularly true for the Cultural Foundation (CF) grades. The correlations of CF grades with all variables are moderate to low. Such a pattern of correlations might indicate the presence of a departmental bias in grades which lacks substantial relationship to current academic criteria being used in other departments. An examination of the means and standard deviations of the grade criteria might also be

TABLE 1

Correlations, Means, and Standard Deviations for MAT and Criterion Measures ($N = 219$)

	Grade Criteria				Masters Comprehensive Exam Scores						
Variables[a]	MSG	CFG	TG	MdTG	EPS	MSS	CFS	TMCS	MAT	$\bar{X}$	S
1. EPG	.37	.25	.48	.46	.40	.45	.39	.48	.31	2.17	.58
2. MSG		.23	.55	.55	.41	.64	.36	.59	.46	2.15	.64
3. CFG			.40	.34	.23	.20	.24	.27	.05	1.91	.59
4. TG				.98	.44	.56	.33	.54	.26	2.32	.29
5. MdTG					.44	.57	.34	.55	.28	2.33[b]	.33
6. EPS						.54	.48	.75	.42	50.41	6.61
7. MSS							.51	.87	.46	59.12	10.82
8. CFS								.77	.44	46.21	8.17
9. TMCS									.51	155.61	21.13
10. MAT										47.32	16.70

[a] Symbols for the variables are explained in the text.
[b] Mean of the median grade point averages.

interpreted as evidencing a ceiling effect, which is associated with lower predictability of the MAT. The correlation of .54 between total grade average and the total scores on the Comprehensive Examination, would indicate that different instructional outcomes are being evaluated. Further evidence bearing on differences in grading practices is found in the moderately low intergrade correlations.

The correlation of .98 between mean and median grade point average would indicate that Chansky's caution is of no practical consequence.

The correlations of .26 between MAT and grades and of .51 between MAT and Comprehensive Examination scores, are typical of those found in the literature. They are interpreted as justifying the continued use of the MAT, although additional predictive validity studies should be periodically undertaken.

REFERENCES

Chansky, N. M. A Note on the Grade Point Average in Research. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 95–99.

Gustad, J. W. A Comparison between the Miller Analogies Test and the Graduate Record Examination as Predictors of Success in Graduate Training. Paper read at Midwest Psychological Association, 1950.

Travers, R. M. W. Unpublished study (1948). Reported in Miller, W. S. *Manual for the Miller Analogies Test,* New York: Psychological Corporation, 1960.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

THE PREDICTIVE VALIDITY OF THE NATIONAL LEAGUE FOR NURSING, PRE-NURSING AND GUIDANCE EXAMINATION FOR DIFFERENT CRITERIA OF SUCCESS IN A THREE YEAR DIPLOMA PROGRAM[1]

GEORGE F. MADAUS
The University of Chicago
Center For the Cooperative Study of Instruction

Problem

This study was undertaken to determine the predictive validity of the National League For Nursing, Pre-Nursing and Guidance Examinations for success in a three year diploma program in a small private hospital school of nursing.

Variables

The National League For Nursing, Pre-Nursing and Guidance Examination (PNLN), a battery made up of different tests of the Cooperative Test Division of Educational Testing Service, is a widely used battery for selection of students in diploma programs. Taylor and others (1963) in their extensive review of selection and recruitment practices in nursing education reported that 288 of 698, or forty one percent, of diploma schools sampled use the PNLN. Forty seven percent of these ranked it as the most important procedure used in selection.

The PNLN covers five areas thought to measure some of "the capabilities and proficiencies which are related to the development of registered nurse competencies and which might reasonably be expected of high school graduates" (National League For Nursing, 1961, p. 3). These are: academic aptitude, reading comprehension,

[1] This work was done in part at the MIT Computation Center, Cambridge Massachusetts.

mathematics, natural sciences, and social studies achievement. These five tests yield the following nine scores: (1) Verbal Aptitude, (2) Quantitative Aptitude, (3) Total Aptitude, (4) Speed of Comprehension, (6) Mathematics Achievement, (7) Natural Science Achievement, (8) Social Studies Achievement and (9) a Composite Score which is a weighted combination of the scores on the individual tests.

Taylor (1963, p. 2) uncovered only three published multiple correlation studies of the predictive validity of the PNLN. With State Boards as the criterion the *R*'s ranged from .61 to .70. When grades were used as the criterion, the obtained *R* was .57.

Criteria

The criteria for success in the three year diploma program were threefold. First, grades in individual courses in the freshman year were used. Unlike reported studies (Taylor, 1963), it was decided neither to lump all courses together nor to utilize an over-all grade point average. This decision was made so that the predictive capacity of the PNLN for clinical and theory courses could be compared, and further to determine the forecasting power of the battery for the different types of science courses, ranging from Pharmacology to Psychology, which the student nurse must master.

Second, the scores of students on four of the Psychological Corporation's Achievement Tests in Nursing were used as criterion measures. These achievement tests, in the areas of Anatomy and Physiology, Chemistry, Microbiology and Pharmacology were taken by students upon completion of course work in each of the areas measured by those tests.

The third and final criterion of success was performance after graduation on State Board Examinations. These tests in the areas of Medical, Surgical, Psychiatric, Pediatric and Maternal nursing comprise the licensing examination for nurses in the Commonwealth of Massachussetts.

Statistical Methods

The data were subjected to an inductive multiple regression analysis.[2] The program which was used instructed the computer to se-

[2] Acknowledgment is made to Dr. Albert Beaton, Jr., of Educational Testing Service for the development of the Computer Program for the IBM 7094, which was used in this study.

lect from the independent variables those which contribute to the multiple correlation in the order of their importance. Under this system, the computer first chooses that variable which correlates most highly with the criterion and then proceeds in order to select the next highest correlation with the effects of the first variable partialed out, and then the next highest multiple correlation with the first and second variables partialed out and so on.

The author tested to see whether the additional variables in the multiple regression equation led to a significant increase in the accuracy of prediction over the correlation obtained with increasingly smaller subsets of the same variables, using the formula for such a test found in McNemar (1960, p. 279).

Whenever the multiple correlation technique was utilized on the PNLN the total aptitude and the composite scores, being combinations of other tests in the battery, were not included in the analysis. Zero order correlations were computed between those two combination scores and the various criterion measures.

Subjects

The subjects were student nurses admitted to a small three year diploma program run by a private hospital in Massachusetts between the years 1961 and 1963.

First year grade point averages were available for 90 students. State Board results and Psychological Corporation's Achievement Test in Nursing scores were available for 36 students.

Results

Table 1 presents the zero order correlation coefficients between the PNLN and the criterion measures of grade point average in various first year nursing courses.

Fifteen of the correlations in Table 1 are significant at the .01 level of confidence, while the remaining sixty six correlations did not differ significantly from a correlation of 0.00.

The Verbal and Quantitative tests add nothing of significance to prediction when used in a multiple regression analysis. The Mathematics and Social Studies tests also add nothing to prediction when the battery is analyzed multivariately. Only one obtained multiple correlation ($R = .30$), between the criterion grade point average in Anatomy and Physiology and the PNLN vari-

TABLE 1

Correlations between National League For Nursing Pre-Nursing and Guidance Examinations and Criterion Measures of Grade Point Average In First Year Nursing Courses ($N = 90$)

Grade Point Average in First Year Nursing Courses	Pre-Nursing and Guidance Examinations (PNLN) Scores								
	Verbal	Quantitative	Speed of Reading Comprehension	Level of Reading Comprehension	Math	Natural Science	Social Science	Total	Composite
Fundamentals of Nursing Course	**	**	.23	**	.32	.44	.28	**	.45
Fundamentals of Nursing Clinical	**	**	**	**	**	**	**	**	**
Medical-Surgical I Course	**	**	**	.23	**	.28	**	**	.31
Medical-Surgical I Clinical	**	**	**	**	**	**	**	**	**
Anatomy and Physiology	**	**	**	**	**	**	**	**	**
Biochemistry	**	**	**	.21	**	.36	.27	**	.36
Microbiology	**	**	.27	.24	**	**	**	**	**
Psychology	**	**	**	.28	**	**	**	**	**
Pharmacology II	**	**	**	**	**	**	**	**	**

** Obtained correlations are not significantly different from a correlation of 0.00

ables of Speed of Comprehension, Level of Comprehension, and Natural Science, was superior in forecasting power to the highest zero order correlation obtained.

When the Psychological Corporation's Achievement Tests in Nursing were used as criterion measures, only one correlation, multiple or zero order was obtained which differed significantly from a correlation of 0.00. This was an r of .37 between the PNLN Mathematics Achievement test and the Microbiology Achievement test of the Achievement Tests in Nursing of Psychological Corporation.

Table 2 shows the zero order correlation coefficients between the PNLN tests and the criteria of performance on State Board Examinations taken three years after administration of the PNLN. The correlations in the last column, pass-fail, are point biserial correlation coefficients; the remaining are zero order correlations.

TABLE 2

Correlation Coefficients between the National League For Nursing Pre-Nursing and Guidance Examinations and Massachusetts State Board Examinations

Pre-Nursing and Guidance Examinations (PNLN)	Massachusetts State Board Examinations					
	Medical	Surgical	Psychiatric	Pediatric	Maternal	Pass-Fail
Speed of Comprehension	.42	.60	**	**	**	.41
Level of Comprehension	**	.55	**	**	**	.45
Mathematics	**	**	.20	.52	.34	.39
Natural Science	**	.34	**	**	**	**
Composite	.45	.62	.42	.33	**	.43

** Obtained Correlations are not Significantly Different from a correlation of 0.00

Fifteen of the 30 correlations in Table 2 are statistically significant at the .05 level of confidence or greater; the remaining fifteen exhibited only chance relationships. Using the seven tests multivariately added nothing in the way of statistically significant increase in prediction over the results obtained in Table 2. The Verbal, Quantitative, Total, and Social Studies tests had no relationship to any of the criteria, and therefore do not appear in Table 2. The Composite score did correlate with all but the Maternal test of the State Boards.

Discussion

The Verbal, Quantitative and Total scores which constitute the academic aptitude section of the PNLN have no statistically significant predictive validity for any of the criterion measures utilized with this sample. Although this is understandable in the clinical areas, in which performance is non-scholastic in nature, the complete lack of relationship with nursing theory and college level science courses is not readily explainable.

None of the tests of the PNLN has predictive validity for clinical performance as measured by grades in courses such as Fundamentals of Nursing Practice, Medical-Surgical Practice, and Pharmacology II. Tests must be found or developed to measure aptitude in these practical, clinical, performance-oriented courses. Behavioral analysis of the clinical duties of registered nurses would be an excellent starting point in the development of such instruments.

The Natural Science and Reading tests appear to be the best predictors of theory courses in nursing and science courses. Although statistically significant, the obtained correlations are of dubious practical value. Multivariate analysis adds little or nothing to prediction. This undoubtedly is due to the relatively-high intercorrelations between the tests in the PNLN battery. A tailor made battery for nursing theory and clinical practice aptitudes should contain tests with low intercorrelations and high individual correlations with the criteria. This is not the case with the PNLN for this sample.

The PNLN did have moderate long range predictive validity when used to forecast performance on the State Board Examinations. The sample was small, thirty-six, and homogeneous because of attrition over the three year span. In spite of this restricted range which tends to suppress correlations, the PNLN Composite score was related to all State Board Tests with the exception of the Maternity Nursing Examination. The Academic Aptitude Tests and the Social Studies Tests had no predictive validity with this criterion, whereas the remaining achievement tests exhibited moderate long-range predictive powers.

Conclusion

Although certain subtests of the PNLN did correlate significantly with various nursing theory and science courses, the relationships were not of an order sufficient for decision making regarding admissions. The crucial area of clinical nursing course performance was unrelated to antecedent performance of any of the tests in the PNLN battery.

Certain Subtests of the PNLN did exhibit long-range predictive validity for this sample for later performance on State Board Examinations.

REFERENCES

McNemar, Quinn, *Psychological Statistics,* New York: John Wiley and Sons, 1960.

Taylor, Calvin W., et al. *Selection and Recruitment of Nurses and Nursing Students.* Salt Lake City: University of Utah Press, 1963.

National League for Nursing. *The N.L.N. Pre-Nursing and Guidance Examination.* Pamphlet No. 1: New York: National League for Nursing, 1961.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

MULTIPLE DISCRIMINANT PREDICTION OF MAJOR FIELD OF STUDY

ROBERT F. STAHMANN and NORMAN E. WALLEN
University of Utah

"Multiple discriminant analysis is a statistical method of combining test scores or other data so as to *maximize* the differences *between* the groups and *minimize* the differences *within* each group (Dunn, 1959, p. 15)." The discriminant analysis, first used as a method of classifying individuals into one of two groups, was introduced into psychological literature in the late 1930's (Travers, 1939). This technique of classification into one of two groups, has been applied to various academic classification problems (Selover, 1942; Baggaley, 1947; Tiedeman and Sternberg, 1952). The discriminant analysis technique has evolved to become a technique which may be used in classifying individuals into one of several groups.

The multiple discriminant analysis has been applied to the problems of predicting academic group membership (Christensen, 1953) and guidance of students based upon such predictions (Stinson, 1958). Dunn (1959) found the discriminant analysis to be superior to the regression analysis for prediction of academic major field. Tatsuoka (1957) applied the multiple discriminant, multiple regression, and "joint-probability model," which was a hybrid of the discriminant and regression analyses, to a sample of students and concluded that for the groups under study the discriminant analysis predicted correctly the greatest number of actual fields of study.

Based upon a multiple discriminant analysis of individual entrance examination data, the present study was an attempt to classify or predict major field of study at graduation for a sample of university students.

Method

University of Utah graduates of 1962, 1963, and 1964, in selected major fields of study served as the population from which the samples were drawn for inclusion in the study. The study required two samples, an experimental sample upon which the discriminant-analysis linear functions were computed, and a cross-validation sample upon which the predictions as to major field of study were made. For the cross-validation sample use was made of the functions found from the discriminant analysis of the experimental sample. A sample size of 50 was desired for each experimental and each cross-validation sample. Random assignment was made to the experimental and cross-validation samples for the fields in which the original number of available subjects exceeded 50. In each of those fields in which the original number of available subjects was less than 50 all subjects for that field were included in the experimental sample. The prediction as to major field of study was run separately for men and women students.

The analysis was run for each problem through using the experimental sample as a normative sample upon which the discriminant weights were computed, and the cross-validation sample upon which the predictions as to major field of study were made. A secondary analysis of the experimental samples was made through classifying each individual in that sample on the basis of the discriminant analysis of the same sample.

The entrance examination battery which made up the data used in the study consisted of the following measures: (1) *Cooperative English Test:* Lower Level, Forms RX, T and Z; (2) *Cooperative Mathematics Pre-Test for College Students,* Forms X and Y; (3) *Cooperative General Achievement Tests: Test 2, Natural Science,* Form YZ; (4) *Occupational Interest Inventory, 1956 Revision,* Advanced level.

Results

The discriminant analysis and classification were facilitated by the use of a standard computer program. The data from the male experimental sample were submitted to the multiple discriminant analysis. The linear discriminant functions for the males were then used in the classification of individuals in the male cross-

validation sample. The discriminant functions obtained from the analysis of the male experimental sample were read into the computer and used as the discriminant weights for predicting major field for the male cross-validation sample. *Each individual was predicted as belonging to the field for which the discrimination score was the highest.* The results of the prediction for the male cross-validation sample are summarized in Table 1.

TABLE 1

Agreement between Predicted and Actual Field of Study, Male Cross-Validation Sample

	Predicted Major Field				
Actual Major Field	Engineering	Business	Pharmacy	Letters & Science	Total
Engineering	29 (17)	13 (10)	3 (8)	5 (15)	50
Business	8 (13)	3 (8)	14 (6)	14 (12)	39
Letters & Science	11 (17)	11 (10)	5 (8)	23 (15)	50
TOTAL	48	27	22	42	139

Note—Top numbers refer to predictions made by the discriminant analysis. Numbers in parentheses are the number of classifications expected by chance, based on marginal totals.

The numbers of correct classifications expected by chance, based on marginal totals, are shown as the numbers in parentheses in Table 1. The frequencies arising from predictions for the male cross-validation sample based upon the weights of the male experimental sample exceed the numbers expected to be correctly predicted by chance for the engineering and the letters and science fields of study. The frequency in the classification for the business field is below the number expected by chance.

The data from the female experimental sample were submitted to the multiple discriminant analysis, and the discriminant weights which were obtained were applied to the predictions for the female cross-validation sample. The prediction as to major field of study for the female cross-validation sample placed 32 of the 50 female elementary education majors in the correct classification. This prediction is significantly better than the chance prediction of 16. The results of the predictions for the female cross-validation sample based upon weights of the female experimental sample are summarized in Table 2.

TABLE 2

Agreement between Predicted and Actual Field of Study, Female Cross-Validation Sample

	Predicted Major Field			
Actual Major Field	Nursing	Elementary Education	Letters & Science	Total
Elementary Education	8 (16)	32 (16)	10 (16)	50

Note—Top numbers refer to predictions made by the discriminant analysis. Numbers in parentheses are the numbers of classifications expected by chance, based on marginal totals.

The computer program which was available for this study was written for use primarily as a classification program setting up the discriminant weights on, and predicting for, the same sample. Such an analysis was run and included in this study. Table 3 summarizes the results of the discriminant analysis of, and prediction for, the male experimental sample. An examination of the results of this prediction reveals that the frequencies associated with predictions for each of the four major fields of study exceed chance expectations.

TABLE 3

Agreement between Predicted and Actual Field of Study, Male Experimental Sample Using Functions of Male Experimental Sample

	Predicted Major Field				
Actual Major Field	Engineering	Business	Pharmacy	Letters & Science	Total
Engineering	38 (15)	9 (13)	2 (8)	1 (14)	50
Business	3 (12)	25 (10)	8 (6)	3 (11)	39
Pharmacy	5 (9)	3 (7)	13 (4)	8 (8)	29
Letters & Science	5 (15)	6 (13)	4 (8)	35 (14)	50
TOTAL	51	43	27	47	168

Note—Top numbers refer to predictions made by the discriminant analysis. Numbers in parentheses are the numbers of classifications expected by chance, based on marginal totals.

A similar analysis, based upon functions of the female experimental sample, and predicting for the members of the same female sample was run. Table 4 summarizes the results of the analysis of, and the prediction for, the female experimental sample. The frequencies associated with predictions for each of the fields of study exceed chance expectations.

TABLE 4

Agreement between Predicted and Actual Field of Study, Female Experimental Sample Using Functions of Female Experimental Sample

	Predicted Major Field			
Actual Major Field	Nursing	Elementary Education	Letters & Science	Total
Nursing	33 (16)	5 (16)	5 (11)	43
Elementary Education	9 (18)	34 (18)	6 (13)	49
Letters & Science	7 (15)	10 (15)	24 (11)	41
TOTAL	49	49	35	133

Note—Top numbers refer to predictions made by the discriminant analysis. Numbers in parentheses are the number of classifications expected by chance, based on marginal totals.

The analysis-of-variance technique was applied to the male experimental and cross-validation samples and to the female experimental sample to examine the differences among the means of the 20 variables for each of the major fields of study. These analyses showed the greatest variability among the means of the interest test rather than among the means of the academic achievement measures.

The t test was applied as a test of equivalence of samples. This analysis revealed that there were no significant differences (.05 level) between the experimental and cross-validation samples on 18 of the 20 variables in any major field.

Summary

Multiple discriminant analysis as a method of predicting in which one of several groups an individual is most likely to place was applied to the problem of predicting major field of study at the University of Utah. Variables used in the study consisted of interest and achievement measures and of an index of urban or rural high school attendance, all of which were obtained from the freshman entrance examination battery of the students.

Generally, the question as to whether the freshman entrance battery at the University of Utah was shown to be an effective predictor of major field of study at graduation can be answered in the affirmative. The prediction for the business field of study was below chance expectation—an outcome tending to support the finding of Christensen (1953) in which prediction at chance was made for

a pre-business group. These findings indicate that in attempting to predict membership in a heterogenous group such as business students, one must perhaps use variables other than achievement and interest measures to achieve adequate discrimination.

REFERENCES

Baggaley, A. R. The Relation between Scores Obtained by Harvard Freshmen on the Kuder Preference Record and Their Fields of Concentration. *Journal of Educational Psychology*, 1947, 38, 421–427.

Christensen, C. M. Multivariate Statistical Analysis of Differences between Pre-Professional Groups of College Students. *Journal of Experimental Education*, 1953, 21, 221–232.

Dunn, F. E. Two Methods for Predicting the Selection of a College Major. *Journal of Counseling Psychology*, 1959, 6, 15–26.

Selover, R. B. A Study of the Sophomore Testing Program at the University of Minnesota. *Journal of Applied Psychology*, 1942, 26, 3 parts, 296–307, 456–467, 587–593.

Stinson, P. J. A Method for Counseling Engineering Students. *Personnel and Guidance Journal*, 1958, 37, 294–295.

Tatsuoka, M. M. Joint-Probability of Membership and Success in a Group: Index Which Combines the Information From Discriminant and Regression Analyses as Applied to the Guidance Problem. *Harvard Studies in Career Development*, No. 6, Graduate School of Education, Harvard University, 1957.

Tiedeman, D. V. and Sternberg, J. J. Information Appropriate for Curriculum Guidance. *Harvard Educational Review*, 1952, 22, 257–274.

Travers, R. M. W. Use of a Discriminant Function in the Treatment of Psychological Group Differences. *Psychometrika*, 1939, 4, 25–32.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

AN ATTEMPT TO VALIDATE AN EMPIRICALLY-DERIVED INTEREST SCALE AND STANDARD KUDER SCALES FOR PREDICTING SUCCESS IN HIGH SCHOOL GEOMETRY

GERALD S. HANNA
University of Alaska

Problem

This paper describes an unsuccessful attempt to measure some of the unaccounted for variance in criteria of learning which, as Jackson and Strattner (1964, p. 513) recently observed, lingers after the effect of ability, prior learning, teaching method, and other task-related variables have been removed. This variance has interested researchers for some time and continues to stimulate research.

Earlier Studies

Super and Crites (1962, pp. 479-81) cited several studies in which significant relationships existed between scales of the Kuder Preference Record (Vocational) and academic success in various fields. Although none of these studies pertained directly to geometry, it appeared likely that similar findings would occur if geometry were investigated. They concluded that the Kuder seems "to have real possibilities even for the prediction of success in courses, for scores are significantly related . . . to grades in some appropriate subjects, specifically the scientific, mathematical, and literary" (Super and Crites, 1962, p. 491).

The only study located which dealt directly with geometry prediction was Townsend's (1945) in which a correlation of .31 was obtained between the Strong Vocational Interest Blank for Men (Mathematician Scale) and the Cooperative Plane Geometry Test

among 63 private secondary school boys. The Mathematics-Science Teacher Scale correlated .28 with the criterion in the same sample.

The Present Study

Standard Kuder Scales

In a recent investigation (Hanna, 1965), numerous predictive variables were obtained for 226 geometry students at the beginning of a school year. Variables included many cognitive measures, past grades in related courses, standard interest scales as well as an empirically-derived interest inventory scale, and pupils' predictions of how well they would achieve in the subject. Criteria included an achievement test for the 202 pupils remaining at the end of the semester, teachers' rankings of students on the semester's work, and a measure of improvement in deductive reasoning in non-geometric situations. Multiple regression equations were derived using selected combinations of predictors and a composite criterion.

Correlations of the ten standard scales of the Kuder, Form CM, with the composite criterion ranged from —.17 to .05; the median absolute value was .08. Only one scale (scientific) correlated significantly (but negatively!) at the .05 level with this criterion. Improvement of multiple regressional prediction by use of existing Kuder scales was equally disappointing. The hope expressed by Super and Crites that existing Kuder keys would prove useful in predicting course success—a use for which the scales were not derived—was not realized in this investigation.

Empirical Scales

A stratified random subsample of 94 Kuder answer sheets was sorted so that top and bottom 27 per cent groups could be separated on the basis of the composite geometry criterion. These groups were then contrasted on every response position for each Kuder item. The ratio used for item discrimination was the difference between the number of responses in the top and bottom groups for a given response position, divided by the average number of responses in that response position for the two groups. Response positions having a discrimination ratio of .4 or more were employed if, and only if, the difference in the actual number of subjects who marked the response position was two or more. For items in which the differ-

ence in the number of pupils who marked the response position was four or more, the item was used in the empirical key, regardless of its discrimination index. Separate keys were made for the positive and negative discriminators. The score for each pupil was determined by subtracting the score obtained when the key of negative discriminating items was used, from the score obtained when the key of positive discriminators was used. The derived keys were then cross-validated with the other 108 subjects by correlating the composite criterion scores with the scores obtained from the empirical keys.

The empirical key correlated .50 with the composite criterion before allowance for shrinkage was made. In cross-validation of this key with the answer sheets of the pupils with which it was not derived, the correlation was only .27. This validity value was used in derivations of multiple regression equations.

Predictors were added by a stepwise regression computer program. In addition to the empirically derived scale, all ten standard Kuder scales were available for computer selection. Successive additions are reported until the next variable would add considerably less than 1 per cent to R^2. Equation (1) reports the multiple regression equation used for a very productive combination of readily obtainable raw data wherein $R = .59$. Equation (2) gives the weights of the same data when used in combination with the standard and empirical Kuder scales; this equation correlated .63 with the composite criterion.

$$\text{(1)} \quad R = .53V + .53W + .50X - .35Y + .09Z$$

$$\text{(2)} \quad R = .52V + .57W + .47X - .40Y + .04Z - .02L + .03M + .09N$$

where,

V = cumulative marks in first year algebra (5 point scale)

W = cumulative marks in last year of general mathematics studied (5 point scale)

X = subjects' own prediction of their geometry marks (5 point scale)

Y = year of high school (4 point scale)

Z = Differential Aptitude Test: Verbal Reasoning plus Numerical Ability raw scores

L = Kuder Mechanical Scale raw score

M = Kuder Persuasive Scale raw score

N = Kuder Social Service Scale raw score

Conclusions

Neither standard nor empirically derived Kuder scales contributed significantly to multiple regression prediction of geometry success, as is evident by their slight weight or absence from equation (2). Apparently, the independent variance which interest factors contributed to multiple regression equations had been tapped about as well as present instrumentation would allow by the more easily-obtained algebra marks, arithmetic marks, subjects' predictions of geometry marks, and year in high school.

It appears from this study that efforts to develop empirical keys to the Kuder Preference Record for geometry prediction would not offer promise.

REFERENCES

Hanna, Gerald S. An Investigation of Selected Ability, Aptitude, Interest, and Personality Characteristics Relevant to Success in High School Geometry. Unpublished doctoral dissertation (University of Southern California, 1965).

Jackson, Philip W. and Nina Strattner. Meaningful Learning and Retention: Noncognitive Variables. *Review of Educational Research*, 1964, 34, 513–29.

Super, Donald E. and John O. Crites. *Appraising Vocational Fitness*, Revised edition (New York: Harper and Row, 1962).

Townsend, A. Achievement and Interest Ratings for Independent School Boys. *Educational Records Bulletin*, 1945, 43, 49–54.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, NO. 2, 1966

VALIDATION OF THREE TESTS OF COGNITIVE STYLE IN VERBALIZATION FOR THE THIRD AND SIXTH GRADES

SARA W. LUNDSTEEN AND WILLIAM B. MICHAEL
University of California, Santa Barbara

ALTHOUGH there exists a history of research in qualitative styles of thinking and verbalization—abstract, functional, and concrete—there are still many research needs in this area. Investigations of qualitative levels of thinking may be traced to Piaget's theoretical work on developmental sequences in concept formation. Russell and Saadeh (1962) summarized the research through 1961 and Lundsteen (1966b) through 1964. From these two articles the following needs for research were implied: (1) the desirability of ascertaining the nature and degree of relationship or congruent validity between experimental measures of qualitative levels of verbalization and other measures purporting to represent verbal abilities—typically standardized tests used in schools; and (2) the development of stimulus material in the form of paragraphs or stories that would tap higher level processes or operations involved in concept formation as well as critical and creative thinking. These operations, in Guilford's (1961) terminology, involve products in a hierarchial order of complexity.

Problem

Accordingly, the major task of this research was to investigate the degree and pattern of interrelationships among three experimental measures intended to reflect three qualitative levels of verbal thinking, a test of critical listening, a standardized test of reading, and a standardized test of verbal ability.

Furthermore, it was hoped that study of performance on the

experimental measures of verbal function at differing age levels would contribute further evidence regarding their validity in relation to other measures of language behavior.

Hypotheses

With the above purposes in mind four major hypotheses were formulated:

1. First, it was hypothesized that if the construct of qualitative levels of thinking—abstract, functional, and concrete—was valid, each of the levels would be correlated highly and positively across all constructed (experimental) measures of the same planned qualitative level. In other words, a high relationship would exist for each of the separate scores for responses designed to be qualitatively different regardless of the stimulus content.

2. Second, there would be a higher relationship between pairs of tests, in which qualitative levels in verbal ability were carefully delineated, than between each of these same tests and other measures of verbal ability without built-in qualitative levels. It was hypothesized that although the relationship of qualitative to non-qualitative measures (as represented by a listening test and two well known standardized tests) would be positive and substantial, it would not be so high that one type of measure could be validly substituted for the other.

3. Third, it was hypothesized that mean differences between scores in tests of words, of paragraphs, and of stories reflecting developmental levels in qualitative thinking at grades three and six would indicate the same preferences (modal preferences) for certain qualitative levels as had been found in earlier studies (Russell and Saadeh, 1962) in which only word stimuli had been used. Specifically, while the third-grade pupils in comparison with the sixth-grade pupils would tend to show significantly higher means for the concrete score, these third-grade pupils would tend to show significantly lower means on the abstract score. Moreover, there would be no significant difference on the average functional score between the two grade levels.

4. Fourth, according to the progressive complexity and length of the stimulus material (word, paragraph, or story) in each of the qualitative measures, there would be differences at each of the two grade levels in the relative degree of preferences for abstract,

functional, and concrete levels of verbalization—i.e., in the cognitive styles demonstrated. It was hypothesized that the more complex materials in the paragraph test and in the story test would be accompanied by an increase in the employment (that is, a regression to) of a concrete cognitive style. (The arbitrary criterion used to define a cognitive style of a pupil was as follows: at least 40 per cent of the responses in the one dominant qualitative category and for each other category at least a per cent of 10 less than that found in the dominant category.)

Definitions

The categories—abstract, functional, and concrete—have been defined elsewhere (Reichard, Schneider and Rapaport, 1944; Russell and Saadeh, 1962; Lundsteen, 1966b). Nevertheless, a brief description of each of these qualities is presented:

Abstract represents a style or quality of verbal expression in which the pupil discriminates by the special features of a class. The child's concept has grown out of common features toward accuracy and breadth of definition. This verbalization is not an individual instance, not isolated, not particular or personal.

Functional characterizes verbal expression using discrimination of objects or situations by use to the child, or by what the child does. The attitude implicit in the verbalization may be even somewhat opportunistic, self-centered. This style is less encompassing, but not quite so particular as the next level. In the earlier version of the word stimulus test (Russell and Saadeh, 1962), the functional category actually included abstract classification involving discrimination by assignment to the operational class of the object, the use or function of the object, not operations of the child himself. This aspect of the wording of responses appeared to effect an undesirable kinship between the abstract and functional categories. Accordingly, the functional responses were revised and constructed in keeping with the pupil-centered definition as stated earlier by Reichard, Schneider, and Rapaport (1944).

Concrete represents a style involving discrimination by nonessential details, incidental features, and isolated particulars often of a highly sensory nature. Individual instances are given as opposed to the distinguishing features of the class. Particulars may involve size, color, time, movement, weight, number, location, and

origin. If the definition involves a "social" concept, i.e., "anger," the child might refer to tone of voice or bodily position. The definition may use the personal words "I" or "my." (The *error* category represented a mistaken idea, completely irrelevant, or misinformation.)

Method

Statistical Treatment

This investigation, classed as a descriptive study of validity, attempted to examine nominal categories and interrelationships between test variables. Through use of correlation programs at the Computing Centers, University of California, Berkeley and Santa Barbara, product-moment correlation coefficients were computed among all possible pairings of variables cited in Table 1. Investigation of development of abilities in qualitative thinking was effected by applying t tests to differences between grade-level means for each of the qualitative scores for each of the three experimental measures. (See Table 2.)

Sample

The subjects for this investigation were 178 children, completing all six tests (to be described) in four third-grade and four sixth-grade classes in Goleta, California. From the parent population, consisting of six schools with 22 classes (678 pupils), the eight classes for the investigation were chosen randomly. The Goleta Union School district, classified as having a typical socioeconomic distribution for suburban schools (Jones, 1965), included 10 elementary schools with kindergarten through sixth grade. The sixth-grade sample appeared to be somewhat representative of the general population of sixth grades because of its close approximation to norms on the reading portion of the Sequential Tests of Educational Progress (STEP Reading) and the School and College Ability Tests (SCAT).

Experimental Measures

To give a general description, a test battery was constructed of variables which would presumably measure levels of thinking (abstract, functional, and concrete) when the responses sought

were stimulated by (1) the word, (2) the paragraph, and (3) the unfinished problem-story. An additional hope was to measure processes of thinking as theorized by Russell (1956).

The instrumentation may be described specifically, measure by measure, as follows:

Concept Formation, Word Test. First, to measure qualitatively the process of concept formation, or definition with a one word stimulus, a multiple-choice test of 42 items with constructed responses for abstract, functional, and concrete qualities was used. This measure, described elsewhere in detail (Russell and Saadeh, 1962; Lundsteen, 1966b), and revised especially for this study, showed a test-retest reliability of .78.

Creative and Critical Paragraph Test. Second, the newly developed test of creative and critical paragraph reading contained in five paragraphs elements of mood, humor, fantasy, and facts. It was hypothesized that the nature of the content made it possible for mental operations of the pupils to approximate Russell's theorized processes of evaluative and creative reading. Or, in Guilford's (1961) terminology, operations involving relations, implications, and transformation were hypothesized as being measured. The 42 words used as stimuli in the test mentioned above were embodied in the five paragraphs in the second test. Examples of questions were inferential types as to character relationships, and critical types, such as, evaluating the mode of presentation as fact or fancy. Forty-five questions, approximately evenly distributed between the paragraphs, dealt with this type of question. Again, as in the first test, an attempt was made to construct the multiple responses in abstract, functional, and concrete styles of language.

Twenty graduate students from the University of California, acting as judges, aided in examining the validity for the response categorization. Test-retest reliability for this measure was .60 ($N = 88$). Percentages of preferences for the three styles at each grade level for each item were also examined.

Problem-Story Test. Third, the experimental test of the problem-solving process developed for this study consisted of 48 questions. There were four different unfinished stories or problems with 12 types of questions accompanying each story. The 12 types of questions, derived from the literature on problem solving, included items grouped under the categories of (1) inference of the problem, (2)

hypotheses for solving the problem, (3) procedures in solving the problem, and (4) evaluation of an hypothesis.

The problem areas for the stories included the following: coping with change, integration into a new peer group, and the rift between the old and the young. When the story reached a point where the problem could be inferred, the selection terminated and the first of twelve question types was asked. This version of the problem-solving test contained the multiple-choice answers representing, again, the abstract, functional, and concrete quality of verbalization.

During two separate item revisions, two classes of 20 to 30 graduate students at the University of California, Santa Barbara, were used in judging the phrasing of the responses on this test. The test-retest reliability was .78 ($N = 88$). Percentages for each quality under each item for each grade level were also inspected.

Each of the three experimental tests was scored by finding for each individual how many items he marked with responses keyed as abstract, functional, and concrete. For example, in the test of concept formation with the word as a stimulus, an individual might have answered 14 responses keyed as abstract, 12 as functional, 8 as concrete, and 8 as errors.

Although all three tests need further revision, the results from the present administration seemed of sufficient interest to report. More measures need to be constructed to represent the process variables.

Additional Measures

Three additional tests were administered to the sixth-grade sample. The test of critical listening has been described elsewhere (Lundsteen, 1963, 1964, 1965, 1966a). This instrument, with an estimated test-retest reliability of .72, measures ability to analyze, interpret, and give reasons for judging a speaker's purpose, propaganda, and argumentation. It was thought that a test of listening ability at higher levels of mental operation would be positively related to the newly constructed measures. Test administration was effected by use of a tape recorder and IBM answer sheets.

Because ascertainment of relationships between the new measures and widely used standardized tests of verbal ability, reading, and achievement was desired, the STEP Reading, Form B and the

SCAT, Form B were administered to the sixth-grade sample. STEP Reading was designed to measure ability to reproduce ideas, to translate ideas, to make inferences, to analyze motivation and presentation, and to criticize. SCAT was developed to measure sentence understanding, numerical computation, word meanings, and numerical problem-solving.

Results

Interrelationships of the Experimental Measures of Qualitative Styles

Inspection of the correlational matrices in Table 1 showed the following results. The first hypothesis proposing that measures of each of the qualities—abstract, functional, and concrete—would be positively and substantially related across the other experimental measures within the (intended) same qualitative level was supported partially. The correlations between abstract subscores ranged from .41 to .72. The functional subscores, however, showed no significant relationship with one another. The correlations varied from .02 to .29. At the third-grade level, the concrete subscores showed insignificant interrelationships. However, there were somewhat stronger interrelationships at the sixth-grade level with one coefficient as high as .50.

Why might these results have taken place? Possibly the reason for the consistently high correspondence of the abstract quality from measure to measure rests in the basic nature of symbolic thinking—a quality or ability perhaps more encompassing and less anchored to the particular stimulus material than are the other two qualities. Accordingly, the variable called abstract quality in thinking—be it in problem solving associated with a story, in critical and creative reading of paragraphs, or in the process of concept formation involving single stimulus words—appears to yield high positive correlations across the experimental measures. This correspondence was especially evident at the sixth-grade stage of development. For concrete responses, perhaps the problem-solving test and the word-meaning test (for which the intercorrelation was .42 for sixth-grade pupils) contained some identical elements in regard to Guilford's products classified as units and classes. Or to phrase the comment another way, perhaps pupils who contributed

TABLE 1

Validity Coefficients for Measures of Qualitative and Non-Qualitative Verbal Abilities

			Intercorrelation of Measures or Subtests			
N	Grade	Test	Total	Abstract	Functional	Concrete
88	3	I & II[a]	.46	.51	.09	.05
		I & III	.45	.41	.03	.32
		II & III	.27	.43	.11	.01
86	6	I & II	.76	.72	.05	.50
		I & III	.62	.56	.29	.42
		II & III	.64	.62	.06	.38
70	6	Critical Listening and—				
		I	.55	.47	.13	−.53
		II	.66	.62	.07	−.47
		III	.67	.63	.02	−.50
70	6	Reading and—				
		I	.58	.50	.11	−.49
		II	.58	.59	−.04	−.48
		III	.66	.56	.20	−.50
70	6	SCAT and—				
		I	.60	.53	.08	−.53
		II	.57	.55	.05	−.47
		III	.67	.59	.13	−.48

[a] I = Problem-story test, 48 items; II = Creative and critical paragraph test, 45 items; III = Concept formation, (word meaning) test, 42 items.

to the concrete scores were operating at these more simple levels irrespective of whether the test afforded products at a more complex level. It should also be remembered that because of the method of scoring these ipsative measures where each answer choice in response to a question represents either the abstract, functional, concrete, or error quality, the magnitude of the correlations may be exaggerated because of experimental dependence in the responses. Consequently the correlational results must be viewed with some caution.

Relationships of the Experimental Measures To Other Measures

The second hypothesis concerning the relationship between the experimental measures and three other tests (critical listening, reading, and scholastic ability) received some support. (See Table 1.) The relationship between the problem-solving story and the creative-critical paragraph tests involving qualitative levels of thinking did indeed show a higher correlation (.76) than did these two measures when each was correlated instead with each of the

two standardized tests, reading and achievement, and with the test of critical listening. The correlations with the non-qualitative measures ranged from .55 to .67. Furthermore, although these correlations are substantial, their magnitude would not necessarily indicate that one measure could meaningfully be substituted for another, even in the case of tests involving problem solving with story content and creative-critical thinking in paragraph stimulus material.

From examination of the intercorrelations of subscores in the abstract, functional, and concrete categories, it was apparent that the main contribution to the positive substantial relationships was coming from the abstract quality. The correlations of the abstract subscores with the three non-qualitative measures varied from .47 to .63 with most of the correlations falling in the .50 to .59 range. It appears that children who read, achieve, and listen well appear to be able to operate with abstract responses.

Developmental Differences

The third hypothesis proposing differences between average scores of third- and sixth-grade pupils in qualities of thinking appeared to be supported. The results were in keeping with the theory stimulated by Piaget and represented in this country by Reichard, Schneider, and Rapaport (1944). Across all three qualitative measures there was significant dominance of the concrete choices for the third-grade children with considerable decline of concrete choices for pupils in the sixth grade. Consistent with the direction of the third general hypothesis, differences between the mean scores on the abstract quality and between mean scores on the concrete quality were statistically significant beyond the .01 level for the pupils in the third and sixth grades. (See Table 2.)

Regarding the functional quality, it appeared that both grades shared this supposedly intermediate step between concrete and abstract thinking qualities. Similar findings were obtained for all three experimental tests.

Impact of the Stimulus Material on Cognitive Style

Examination of the data in regard to the fourth hypothesis gave some surprising results. Individual pupils were classified with respect to the predominant quality of their performance on each of

TABLE 2

Significance of Differences between Test Scores Reflecting Categories of Choices Chosen as "Best" by 174 Children

Test	Category	Grade 3 $N = 88$ M	Grade 3 $N = 88$ SD	Grade 6 $N = 86$ M	Grade 6 $N = 86$ SD	t values
1. Problem-Story						
	Abstr.	10.77	4.08	19.01	5.28	11.52*
	Funct.	10.90	2.65	11.65	3.04	1.74
	Concr.	11.67	3.79	6.20	3.74	9.55*
2. Creative & Critical Paragraph						
	Abstr.	14.06	5.69	21.22	6.40	7.80*
	Funct.	12.42	3.73	11.57	3.27	1.60
	Concr.	15.79	21.57	8.80	3.95	2.96*
3. Concept Formation (Word Meaning)						
	Abstr.	14.75	5.78	23.25	6.09	9.49*
	Funct.	16.02	3.75	16.52	3.70	.88
	Concr.	11.54	4.39	5.64	3.52	9.78*

* Significant at .01 level, $t = 2.60$.

the three experimental tests. Each child's dominant preference for any one quality—abstract, functional, or concrete—was identified as mentioned previously by an arbitrary criterion of at least 40 per cent of total number of responses in that category and at least 10 per cent fewer of the total responses in any other category. The following results are given in percentages for 90 third-grade pupils and 88 sixth-grade pupils. Although only 24 per cent of the third-grade pupils exhibited a distinct cognitive style, the majority of this group demonstrated a concrete style on the word sitmulus test (13%). Nine per cent of the third grade pupils displayed an abstract style and two per cent displayed a functional style. Of those showing a dominant cognitive mode in the sixth grade, the majority exhibited an abstract style (67%). The rest did not reach the criterion. A higher proportion of sixth-grade pupils than of third-grade pupils did exhibit a style.

Regarding the more complex materials, the story and the paragraph, the sixth-grade exhibition of style remained much the same. For the third grade, however, there was an interesting and surprising trend. Instead of an increase in concrete cognitive style pre-

dicted for the more complex stimulus materials, third-grade pupils exhibited an increase in abstract style. As the stimulus materials became more lengthy and complex in the paragraph and in the problem-solving story units, as the questions supposedly provoked more complex operations, it appeared that more, rather than fewer, third-grade pupils were tending to think with an abstract style. Perhaps the major factor was that the latter two experimental measures (problem-story test and creative and critical paragraph test) were simply more intriguing and interesting than was the word stimulus test. For the concept formation (word stimulus) test 9 per cent of the third-grade styles fell into the concrete category. For the problem-story and the critical and creative paragraph test the percentage of abstract styles, at the third-grade level, increased to 21 per cent. An ingredient of affective involvement stimulated by the more challenging material may have brought about, in part, this more than doubling of percentages, that is, from 9 per cent to 21 per cent.

Results from a Factor Analysis

Finally, although the details of the factor analysis are not presented because of the limitations posed by experimental dependence among the measures associated with ipsative properties in scoring, a few general findings suggested some possible noteworthy relationships consistent with the previous correlational findings.

Although the variables associated with subscores on concrete qualities from the problem-story test and the paragraph test appeared on the same factor at sixth-grade level, the concrete subscore from the concept formation (word stimulus) test was not related to this factor. In effect, all functional variables and all concrete variables did not appear on one functional factor or on one concrete factor at the sixth-grade level, but two of the more complex concrete variables were related to the same factor. The relationship of these two complex concrete variables is consonant with the assumption that creative and critical thinking processes are interrelated within the broad process of problem solving.

At the third-grade level, however, there appeared to be no interrelation among operations at the concrete level. Indeed, there was a unique factor for each concrete subscore category in each of the experimental tests. However, loadings were present for two ab-

stract variables on a common factor at the third-grade level. Specifically, both the problem-solving story test and the creative-critical paragraph test showed loadings on the same factor. From these results it may be inferred that if a third-grade pupil is capable of abstract thinking, his pattern of interrelationship might more closely approximate the factor structure of sixth-grade pupils or more mature students.

Finally, with respect to the factor of concrete concept formation, the test measures—critical listening, STEP Reading, and SCAT—showed high negative loadings. This factorial result is consonant with the correlational findings. Stated negatively, it appears that children who tend to exhibit concrete qualities of concept formation do not do well on tests of critical listening, reading, and scholastic achievement.

Conclusion

A study was made to ascertain interrelationships at each of two grade levels between three experimental tests purporting to measure three qualitative levels of verbal functioning, a test of critical listening, STEP Reading, and SCAT. A theoretical framework stemming from early work by Piaget and set forth by Russell was the major basis for construction and validation of the tests.

Results indicated that measures of verbal thinking with an abstract quality are closely related irrespective of whether the stimulus material is a word, a paragraph, or a story. Relationships, however, do not appear high enough to justify the substitution of one process measure for another. Functional and concrete qualities of thinking are not related significantly either to the measure of listening or to standardized tests of reading and general verbal ability used in the present investigation. Developmentally, correlational data for each of the new or revised measures seem to sustain the theoretical framework from earlier investigations. There was a significant difference (.01 level) between the means for the third- and sixth-grade group on the abstract and concrete scores for all three measures. Irrespective of which of three experimental tests was employed, the third-grade means were substantially higher on the measures of the concrete quality than on the measures of the other two qualities. In contrast, sixth-grade means were noticeably higher on the measures of abstract quality than on the concrete or

functional qualities. Lastly, increase in complexity of verbal stimulus material, however, appeared to stimulate a preference for abstract cognitive style even at the third-grade level. The appearance of this increase should be provocative of further research in test development at the third-grade level, in which specific training and carefully devised materials might be used to stimulate additional abstract thinking in younger children.

REFERENCES

Guilford, J. P. Factorial Angles to Psychology. *Psychological Review,* 1961, 68, 1–20.

Jones, Jack B. Influence of Professional Mothers on Reading Achievement of Sixth-Grade Children. Unpublished master's thesis, University of California, Santa Barbara, 1965.

Lundsteen, Sara W. Teaching Abilities in Critical Listening in the Fifth and Sixth Grades. Unpublished doctoral dissertation, University of California, Berkeley, 1963.

Lundsteen, Sara W. Teaching and Testing Critical Listening in the Fifth and Sixth Grades. *Elementary English,* 1964, 41, 743–747.

Lundsteen, Sara W. Critical Listening—Permanency and Transfer of Gains Made During an Experiment in the Fifth and Sixth Grades. *California Journal of Educational Research,* 1965, 16, 210–216.

Lundsteen, Sara W. Teaching and Testing Critical Listening, An Experiment. *Elementary School Journal.* 1966, 66, 311–315. (a)

Lundsteen, Sara W. Qualitative Levels in Children's Thinking. Paper read at the annual meeting of the AERA, February, 1966. (b)

Reichard, S., Schneider, M., and Rapaport, D. The Development of Concept Formation in Children. *American Journal of Orthopsychiatry,* 1944, 14, 156–161.

Russell, David H. *Children's Thinking.* Boston, Massachusetts: Ginn and Company, 1956.

Russell, David H. and Saadeh, Ibrahim Q. Qualitative Levels in Children's Vocabularies. *Journal of Educational Psychology,* 1962, 53, 170–174.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

A FOURTH VALIDATION OF A READING PROGNOSIS TEST FOR CHILDREN OF VARYING SOCIO-ECONOMIC STATUS

SHIRLEY FELDMANN[1]
The City College
AND
MAX WEINER[1]
Brooklyn College

THE Reading Prognosis Test (RPT) was constructed to provide adequate measures of skills underlying reading of children in both middle and lower socio-economic levels. It consists of seven subtests, grouped in three areas.

In the Beginning Reading area are tests of Alphabet Letters and Sight Vocabulary. The Perceptual Discrimination area includes tests of Auditory Discrimination, Visual Similarities, and Visual Discrimination. In the Language area are tests of Meaning Vocabulary and Storytelling.

Previous validation studies with the RPT indicated that it was a good overall predictor of reading achievement as measured by standardized reading achievement tests administered at the end of the first or second grade (Weiner and Feldmann, 1963). The present study was undertaken during the school year 1963–1964 in order to validate a revised form of the RPT.

Test Item Changes from Third to Fourth Study

Since the third validation study indicated a need to revise certain items and directions, the subsequent revision of the RPT included the changes described below.

[1] Consultants: Institute for Developmental Studies, Department of Psychiatry, New York Medical College.

First, a new Meaning Vocabulary subtest was constructed through using a completely different set of words. The new words were selected on the basis of frequency of use by different grade levels as listed in the Rinsland Word List (1945), and because of their statistical comparability to the words formerly used.

Next, because in the three previous studies the items of the Auditory Discrimination Test showed little discrimination power, a list of 26 items, most of them new items consisting mainly of nonsense words, was introduced. Another problem with the Auditory Discrimination Test was in the clarity of its directions. They were changed to include a clearer statement to the respondent of what is meant by "same" and "not the same," a concept which children of lower socio-economic level have found difficult. This subtest was retained despite its difficulties because previous studies at the Institute for Developmental Studies had found positive results with an Auditory Discrimination Test in relation to reading achievement (Katz and Deutsch, 1963).

The remaining subtests, Sight Vocabulary, Alphabet Letters, Storytelling, Visual Discrimination, Visual Similarities, and Beginning Reading, did not require revision. However, general directions for all tests, directions to be read to the pupils, and scoring procedures were revised.

Procedures

For the present study, a new form of the Institute for Developmental Studies' Socio-Economic Classification Scale was used to determine Socio-Economic Status (SES) of the subjects. In the new scale the children were classified into three SES levels, with the low group as SES I and the high group as SES III. Only two classifications had been used in earlier studies.

One advantage of the three-level categorization is that the difference between Levels I and III seems more definitive than previously.

The subjects were both drawn from selected schools in New York City and included the entire first grade population in a small Westchester community. All the children were tested with the RPT during the first six weeks of the first grade. At the end of the school year, either the Gates Primary Reading Tests (Paragraph Reading) or two parts of the Metropolitan Achievement Tests Primary

I Battery Word Knowledge, and Primary I Reading were administered to the subjects.

Tables 1–6 report the intercorrelation coefficients among Reading Prognosis subtest area, total test scores, and criterion scores, grouped according to socio-economic levels and community. As one might expect, the Beginning Reading area appears to be the best predictor of end-of-first-grade reading ability for all groups. The Perceptual Discrimination area is the next best predictor, whereas the Language area yields the lowest r's with the criterion measures.

It appears that the Language and Perceptual Discrimination areas are less powerful predictors of reading achievement because of the presence of one subtest within each area respectively, Storytelling in the Language area and Auditory Discrimination in the Perceptual Discrimination area. Further analysis of the Storytelling scores revealed that of the 21 subjects who earned a raw score of 3 or less, 16 earned a 1.61 or lower grade equivalent on the Metropolitan Achievement Primary I Reading Test. Analysis of the Auditory Discrimination subtest showed that of those 30 children who earned raw scores below 8, 18 earned a grade score of 1.6 or lower on the criterion achievement test.

For the sample for whom the Gates Primary Reading Tests Paragraph Reading served as the criterion measure, the same relationship was found between low scores on the two RPT subtests and low scores on the criterion test. Although high scores on these subtests may be rather meaningless, low scores may prove useful for diagnostic purposes.

Discussion

It is evident from Tables 1–6, that the total score of the RPT is a good predictor of reading as measured by two different standardized reading achievement tests for children from different communities and different socio-economic levels. Total scores obtained at the end of kindergarten or at the beginning of the first grade would appear to be useful for purposes of grouping for instruction as well as for anticipating future reading levels.

The area scores, Beginning Reading, Perceptual Discrimination, and Language, are also good predictors of reading achievement. In spite of some problems with two subtests, Auditory Discrimina-

TABLE 1

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores
(Community A, SES I, N = 36)

	BR	L	PD	TOT	MATWK RS	MATWK GS	MATR RS	MATR GS
1. Beginning Reading (BR)		.539	.760	.930	.651	.603	.725	.690
2. Language (L)			.351	.675	.416	.411	.344	.349
3. Perceptual Discrimination (PD)				.889	.539	.493	.508	.481
4. Total Score (TOT)					.643	.600	.641	.616
5. Metropolitan Achievement Test (MATWKRS) Primary I, Word Knowledge-Raw Score						.981	.614	.597
6. Metropolitan Achievement Test (MATWKGS) Primary I, Word Knowledge-Grade Score							.516	.556
7. Metropolitan Achievement Test (MATRS) Primary I, Reading-Raw Score								.990
8. Metropolitan Achievement Test (MATGS) Primary I, Reading-Grade Score								
Mean	5.19	9.69	20.14	34.97	14.00	1.50	14.89	1.58
S.D.	5.84	3.84	6.28	13.61	7.02	.25	7.18	.30

TABLE 2

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores
(Community A, SES II, N = 33)

	BR	L	PD	TOT	MATWK RS	MATWK GS	MATR RS	MATR GS
1. Beginning Reading (BR)		.762	.586	.923	.806	.855	.584	.610
2. Language (L)			.368	.780	.551	.668	.430	.428
3. Perceptual Discrimination (PD)				.807	.699	.675	.590	.600
4. Total Score (TOT)					.825	.871	.645	.659
5. Metropolitan Achievement Test (MATWKRS) Primary I, Word Knowledge-Raw Score						.930	.658	.554
6. Metropolitan Achievement Test (MATWKGS) Primary I, Word Knowledge-Grade Score							.670	.563
7. Metropolitan Achievement Test (MATRRS) Primary I, Reading-Raw Score								.639
8. Metropolitan Achievement Test (MATRGS) Primary I, Reading-Grade Score								
Mean	7.15	11.67	22.12	40.94	19.09	1.75	18.15	1.76
S.D.	5.67	4.11	5.86	13.25	7.78	.40	7.93	.39

TABLE 3

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores
(Community A, SES III, N = 109)

	BR	L	PD	TOT	MATWK RS	MATWK GS	MATR RS	MATR GS
1. Beginning Reading (BR)		.348	.464	.831	.611	.618	.643	.660
2. Language (L)			.429	.690	.397	.309	.289	.304
3. Perceptual Discrimination (PD)				.806	.486	.419	.577	.530
4. Total Score (TOT)					.654	.600	.674	.668
5. Metropolitan Achievement Test (MATWKRS) Primary I, Word Knowledge-Raw Score						.902	.801	.741
6. Metropolitan Achievement Test (MATWKGS) Primary I, Word Knowledge-Grade Score							.793	.804
7. Metropolitan Achievement Test (MATRRS) Primary I, Reading-Raw Score								.944
8. Metropolitan Achievement Test (MATRGS) Primary I, Reading-Grade Score								
Mean	11.70	13.93	26.45	52.07	26.93	2.19	28.26	2.24
S.D.	6.32	3.97	5.15	12.14	6.86	.51	11.39	.67

TABLE 4

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores (Community B, SES I, N = 19)

	BR	L	PD	TOT	GPPR RS	GPPR GS
1. Beginning Reading (BR)		.340	.704	.879	.817	.816
2. Language (L)			.246	.560	.294	.322
3. Perceptual Discrimination (PD)				.896	.738	.739
4. Total Score (TOT)					.811	.820
5. Gates Primary Reading Tests (GPPRRS) Paragraph Reading-Raw Score						.992
6. Gates Primary Reading Tests (GPPRGS) Paragraph Reading-Grade Score						
Mean	4.11	8.00	15.31	27.42	6.36	1.75
S.D.	3.74	2.77	5.18	9.48	5.87	.45

TABLE 5

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores (Community B, SES II, N = 11)

	BR	L	PD	TOT	GPPR RS	GPPR GS
1. Beginning Reading (BR)		.123	.760	.885	.750	.638
2. Language (L)			.539	.523	−.144	−.308
3. Perceptual Discrimination (PD)				.956	.573	.361
4. Total Score (TOT)					.612	.427
5. Gates Primary Reading Tests (GPPRRS) Paragraph Reading-Raw Score						.935
6. Gates Primary Reading Tests (GPPRGS) Paragraph Reading-Grade Score						
Mean	13.54	12.63	23.18	49.27	16.73	2.59
S.D.	5.87	2.42	5.39	11.63	5.99	.58

tion and Storytelling, subtest scores as well as total area score should be valuable as diagnostic measures of pre-reading skills.

The present study also indicated that another value of the RPT test is its individual administration. The teachers who had administered the test at the beginning of the school year had an opportunity to learn how their pupils respond to certain questions

TABLE 6

Intercorrelations of Reading Prognosis Subtest Area, Total Test, and Criterion Test Scores (Community B, SES III, N = 18)

	BR	L	PD	TOT	GPPR RS	GPPR GS
1. Beginning Reading (BR)		.443	.780	.900	.712	.659
2. Language (L)			.442	.719	.515	.412
3. Perceptual Discrimination (PD)				.895	.800	.666
4. Total Score (TOT)					.810	.696
5. Gates Primary Reading Tests (GPPRRS) Paragraph Reading-Raw Score						.922
6. Gates Primary Reading Tests (GPPRGS) Paragraph Reading-Grade Score						
Mean	14.72	12.11	24.00	50.83	15.00	2.55
S.D.	5.69	4.65	5.43	13.33	7.00	.85

in areas related to reading. Determination of underlying skills with a standardized procedure should add to the teacher's understanding of each pupil as well as to the teacher's implementation of the curriculum.

The RPT used in the present validation study is still undergoing minor revisions based on item analyses of the subtests. After pilot studies of these revisions, the test will be ready for standardization.

REFERENCES

Katz, P. and Deutsch, M. Visual and Auditory Efficiency and its Relationship to Reading in Children. *Cooperative Research Project No. 1099, Report*, 1963.

Rinsland, Henry. *A Basic Vocabulary of Elementary School Children.* New York: Macmillan Co., 1945.

Weiner, M. and Feldmann, S. Validation Studies of a Reading Prognosis Test for Children of Lower and Middle Socio-Economic Status. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 807–814.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

CONSTRUCT VALIDITY OF DUNCAN'S PERSONALITY INTEGRATION SCALE

LOGAN WRIGHT
Purdue University

THE present investigation was designed as a construct validity study of the Duncan Personality Integration Scale (1963). Recent developments in clinical psychology and personality theory have emphasized the importance of conceptualizing and investigating integrated behavior or positive mental health (Wishner, 1955; Shoben, 1957; Seeman, 1963). This emphasis has created a strong need for objective and quantifiable instruments to measure such behavior. Tests by Duncan (1963) and Fitts (1965) are apparently the only psychometric devices which are available for this purpose. These tests require immediate investigation of their validity so that they might either be employed to assess integrated behavior in both clinical and research situations, or be discarded in order that the search for more adequate measures may be begun.

Items in the Duncan Scale derive content validity by virtue of the fact that they sample the six types of behavior which, according to a review by Jahoda (1958) of personality theory and research in the area, encompass the area of positive mental health. Specifically, the Duncan scale was designed to assist in researching Seeman's (1959) theory of organismic integration.

This instrument (see Table 1) requires subjects to nominate peers as being one of the three members of their group who is best described by a given item, such as: "Who are the three persons who seem best able to deal effectively with everyday tensions and anxieties?" Early attempts at construct validation of the scale have produced positive results. For instance, in one study (Duncan, 1964), integrated subjects, in contrast to those who were less well

integrated, showed more positive self concepts, greater internal locus of control, more extra-curricular involvement, and a higher grade point average even though they did not possess high aptitude scores. This additional validity study sought to examine the relationship of the Duncan Scale to response styles, to self concept, and to the reported perception of others as measured by a sentence completion technique described by Getzels (1951).

The sentence completion task was a paired direct-projective questionnaire (PGPQ) consisting of two sets of 95 open-ended stems, which differed only in that one set was presented in the first person and the other in third person form. Seventeen of the 95 items were personal in nature, e.g., "after I've finished talking, I usually feel that. . . ." The remaining 78 items were political items, e.g., "Helen feels that governmental rent controls should be . . . ," which served as fillers.

Getzels (1951) hypothesized that when subjects respond to third person items under speeded administration conditions, they tend to express personal feelings, since these are readily available to them. The third person form supposedly prevents subjects from recognizing that personal feelings are being expressed, and thus elicits material which, compared with that in the first person, is less socially desirable and less a part of conscious experience. Responses to first person stems are felt by Getzels to reflect feelings which are more socially desirable and which are more a part of conscious experience than are those responses to third person stems. Theoretically, then, the 17 personal items, when completed in both first and third forms, should provide measures of self concept (positivism or negativism in first person responses), reported perception of others (positivism or negativism in third person responses), congruence of responses (presence or absence and magnitude of discrepancies between first and third person responses for the same item), and degree of commitment (Rubin-Rabson, 1954) in response set (the extent or degree of positivism or negativism employed in responding to a PDPQ item).

Method

Subjects

The subjects were 49 single white female undergraduates at a large midwestern university. Each subject resided in one of 4 units

which housed 14 to 16 sorority members. Subjects had been allowed to choose a roommate; however, mates were randomly assigned to units.

Procedure

Testing was done in two one-hour sessions which were separated by one week. During the first session, subjects were administered Form A of the PDPQ task. This form presented seven of the personal items and 38 of the fillers in first person form, with the remaining personal and filler items given in the third person. The following instructions were to read to provide a speeded condition:

> This is a test of how fast you can think. Complete each of the following sentences so that it makes the best sense possible. You may use either a word or a phrase, although a phrase is preferable. Since you will have only a limited time, work quickly. In most cases the best way to answer the test is to put down the first thing that comes to your mind. Example: The unfinished sentence: When Ann scored the goal, she . . . may be finished by adding "felt good" or "won the game" or "told her friend" or "wanted to score another" or something like that. This is a speed test. Work fast. Do not skip any sentences.

At the second session, subjects completed Form B of the PDPQ which consisted of the same 95 items as Form A, but with the person of each stem switched from first to third or vice versa. The same speeded instructions utilized with Form A were employed. After completing Form B, subjects rated the other residents of their own housing unit by means of the Duncan Scale.

Six of the subjects who were present at the first testing did not participate in the second session. Statistical analyses were performed on the data provided by the remaining 43 subjects who completed the entire battery. Scores on the Duncan Scale were corrected to allow for the number of residents from a given unit participating in the study, since the subjects' scores depend upon the number of other subjects rating them.

Each of the PDPQ responses was scored for positivism-negativism by two raters on a 1–5 Likert-type scale. The raters were naive in terms of previous PDPQ rating experience and also as far as knowledge of subjects' scores on the integration scale. Inter-rater reliability was estimated by a Spearman rank-order correla-

tion between the two raters' scores for the 95 first person and 95 third person items provided by a single subject. Each of the Duncan Scale items, as well as the total score, was correlated by means of a Pearson product-moment solution with each of the following three PDPQ variables: positiveness of all first person responses, positiveness of all third person responses, and sum of discrepancies between all first and third person ratings.

In order to determine extent of commitment in response set, a Pearson product-moment correlation was estimated between integration scores and the extent to which subjects' responses for all items deviated from the neutral central tendency on the 1–5 Likert scale for positivism-negativism.

Results

Inter-rater reliability coefficient for the PDPQ was estimated to be .92. Correlations between the Duncan Scale items and the three PDPQ variables are shown in Table 1. Ten of these correlations were significant at a level less than .05. Only one such correlation would be expected by chance.

As predicted, the Duncan Scale correlated with positiveness of the first person ratings on the PDPQ. Three of the six scale items, as well as the total score, correlated positively and significantly. The other three integration items correlated positively though not significantly.

Also, as predicted, the Duncan Scale correlated with positiveness of third person ratings on the PDPQ. Two of the items, in addition to the total score, correlated positively and significantly with this variable. The other four items correlated positively though not significantly with this variable.

Contrary to prediction, item three of the Duncan Scale correlated significantly with the sum of all PDPQ first and third person discrepancies. Also, contrary to prediction, all five remaining items correlated positively, though not significantly with this variable.

It will be noted that all six integration variables correlated in a positive direction with each of the three PDPQ variables. The probability of all six integration variables correlating in the same direction with a variable by chance, as measured by a two tailed sign test (Siegel, 1956, p. 68) is .032.

The correlation between subjects' integration scores and their

TABLE 1

Correlation of Seven Duncan Scale Items with Three PDPQ Variables

Duncan Scale Items	Positiveness of all First Person Responses	Positiveness of all Third Person Responses	Sum of Discrepancies Between all First and Third Person Ratings
1. Who are the three persons in your unit who seem best able to express their feelings without hurting the feelings of others?	.18	.35*	.14
2. In your opinion who are the three persons in your unit who seem to understand themselves best; that is, are aware of their shortcomings and strengths?	.33*	.34*	.21
3. Who are the ones in your unit who seem best able to keep an open mind and not jump to premature conclusions?	.34*	.13	.37*
4. Who are the three persons in your unit who seem the most able to deal effectively with everyday tensions and anxieties?	.25	.14	.09
5. Which three persons in your unit seem capable of forming deeper and more profound relationships with others and seem to be genuinely concerned with other people?	.11	.11	.03
6. Which three persons in your unit seem to you to have been the most successful in all phases of their life: social, personal, educational, etc.?	.31*	.24	.14
7. Total Score	.33*	.35*	.21

$^{*} p < .05$

tendency to make positive or negative responses which deviate from the neutral central tendency was designed to relate integration to commitment in response set. This correlation was .38, which is significant at the .05 level.

Discussion

Performance on the Duncan Scale appears to be related to a positive self concept. This notion is supported by the fact that three Duncan scale items, as well as the total score, correlated positively and significantly with the positive quality of the first person ratings

on the PDPQ. These data support earlier findings and contribute to the construct validity of the Duncan Scale.

Ratings on the Duncan Scale were also related to a more positively reported perception of others. This is indicated by the fact that two of the integration items, as well as the total score, correlated positively and significantly with the positive quality of the third person ratings. This is consistent with Rogerian self theory which assumes that the person who accepts himself will be, because of this self-acceptance, more accepting of others (Rogers, 1951, p. 520). This finding is also felt to contribute to the construct validity of the Duncan Scale.

There was a positive relationship between the Duncan Scale and the sum of discrepancies on the PDPQ. If this result is interpreted according to Getzels' hypothesis, it suggests that integrated subjects are less congruent in their response style than are nonintegrated subjects. This conclusion coupled with results of a study by Getzels (1951) which utilized normal and disturbed subjects would suggest that congruence, as measured by a PDPQ type instrument, is curvilinearly related to integrated behavior. However, the fact that persons scoring high on the Duncan Scale tended to have more discrepancies in first and third person responses on the PDPQ can be explained in part by their tendency to make more committed responses. Subjects scoring low on the Duncan Scale made more neutral responses, and neutral responses produce fewer discrepancies. Also, the discrepancies of subjects who scored high on the Duncan Scale were not the result of their completing the first person items with positive statements and the third person items with negative statements as Getzels' disturbed subjects had done. The higher sum of discrepancies for integrated subjects resulted from the fact that they possessed the greater tendency not only to rate themselves higher than others, but also to rate others higher than themselves. In this case, the word "openness" seems to describe better their behavior.

This finding raises the possibility that integrated subjects were less likely than disturbed subjects to project personal feelings in responding to third person stems. If so, the third person stems, in contrast to first person stems, elicit responses which are more indicative of the integrated subjects' true perceptions of others. If this hypothesis is true, it points up a methodological problem that must

be considered in research on personality integration: constructs may need to be operationally defined in a manner for effective subjects different from that for normal or disturbed ones.

Summary

Forty-three single white female undergraduate sorority members at a large midwestern university were administered the Duncan Scale, a measure of personality integration, and a projective-direct personality questionnaire (PDPQ). The latter scale yielded scores for commitment and congruence in response styles, self concept, and reported perceptions of others. Ten of the 21 correlations between the integration scale and PDPQ variables were significant. It was concluded that performance on the Duncan Scale is related to a relatively high positive concept of both self and others, as well as to a tendency to make committed, as opposed to non-commital, responses. These findings were interpreted as supporting the construct validity of the Duncan Scale. Reference was made to a methodological problem concerning the operational definition of constructs for effective subjects.

REFERENCES

Duncan, Cary. Duncan Personality Integration Scale. Unpublished Test, Nashville, Tenn., George Peabody College, 1963.

Duncan, Cary. Construct Validity of a Reputation Test of Personality Integration. Unpublished doctoral dissertation, Nashville, Tennessee, George Peabody College, 1964.

Fitts, W. H. *Tennessee Self Concept Scale,* Counselor Recordings and Tests, Nashville, 1965.

Getzels, J. W. The Assessment of Personality and Prejudice by the Method of Paired Direct and Projective Questions. Unpublished doctoral dissertation, Cambridge, Massachusetts, Harvard, 1951.

Jahoda, Marie. *Current Concepts of Positive Mental Health.* New York, Basic Books Inc., 1958.

Rogers, C. R. *Client Centered Therapy.* Boston: Houghton Mifflin, 1951.

Rubin-Rabson, G. Correlates of the Non-Committal Test-Item Response. *Journal of Clinical Psychology,* 1954, 10, 93–95.

Seeman, Julius. Toward a Concept of Personality Integration. *American Psychologist,* 1959, 14, 633–637.

Seeman, Julius. Studies in Personality Integration. *Peabody Papers in Human Development,* 1963, 1, 1–9.

Shoben, E. J. Toward a Concept of Normal Personality. *American Psychologist,* 1957, 12, 183–189.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

Wishner, Julius. The Concept of Efficiency in Psychological Health and in Psychopathology. *Psychological Review,* 1955, 62, 69–80.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

IDENTIFICATION OF FOUR ENVIRONMENTAL PRESS FACTORS IN THE STERN HIGH SCHOOL CHARACTERISTICS INDEX

HOWARD R. KIGHT

AND

EDWIN L. HERR

State University of New York, Buffalo

Problem

The purpose of this investigation was to ascertain the factored dimensions of the 30 subscales on the Stern High School Characteristics Index (HSCI). The HSCI is one of four instruments developed by Stern, Stein, and Bloom (1956) to measure environmental press variables but is the only one specifically designed for the high school environment. Wolf (1965) has recently called attention to the growing importance of assessments of this type and the need for validating research evidence.

Procedure

Two samples of 725 and 348 students from two different high schools were administered the HSCI. A principal components analysis was performed from use of the 30 subscale scores obtained for each sample and then rotated analytically according to Kaiser's (1960) varimax criterion.

Results

Although six factors were extracted with the first sample (A) and five factors from the second (B), only the first four factors in each sample, which appeared to be invariant, could be interpreted meaningfully. They were identified as:

1. *Social-intellectual avoidance*—a tendency to avoid or withdraw from intellectual and/or social situations which might be competitive or lead to frustration, humiliation, and a loss of self-esteem. (Factor I)

2. *Inferiority reaction*—a tendency to counteract criticism or feelings of inferiority by becoming indifferent to or disregarding the opinions and feelings of others. May also be manifested into overt forms of behavior. (Factor II)

TABLE 1

Factor Loadings of Thirty Subscales on the Stern High School Characteristics Index For Two Samples
(N_a = 725, N_b = 348)

	Factors							
	I		II		III		IV	
Subscales	A	B	A	B	A	B	A	B
1. Abasement	25	34	73	72	−05	04	22	−05
2. Achievement	−60	−56	−14	−15	20	06	−20	20
3. Adaptability	04	07	69	65	10	11	02	29
4. Affiliation	−40	−67	−26	−20	03	−09	−62	26
5. Aggression	17	20	64	71	−38	−25	12	07
6. Change	01	−31	01	22	−11	−29	−09	−11
7. Conjunctivity	−43	−49	−32	−49	37	10	−32	30
8. Counteraction	−26	−59	−13	−10	15	00	−06	06
9. Deference	−04	04	01	−17	60	23	02	04
10. Dominance	10	04	58	49	−04	−09	−23	56
11. Ego-Achievement	−63	−75	−04	−08	15	−05	−21	−06
12. Emotionality	−18	−26	04	04	−04	−44	−03	26
13. Energy	−58	−63	−25	−28	18	−08	−32	17
14. Exhibitionism	−47	−63	04	−13	05	−12	−53	39
15. Fantasied Achievement	−51	−24	07	11	−17	−02	21	−04
16. Harm Avoidance	−32	−22	−36	−20	42	25	08	−04
17. Humanism	−76	−75	−09	−10	13	19	−04	−07
18. Impulsiveness	04	22	23	−01	−38	−70	−17	20
19. Narcissism	−29	−47	−03	−28	29	13	−44	53
20. Nurturance	−59	−60	−19	−25	23	15	−26	13
21. Objectivity	−35	−32	−68	−77	14	−01	−16	01
22. Order	−22	−21	−03	−11	64	49	−17	44
23. Play	−15	−42	−13	−22	−13	−38	−74	36
24. Practicalness	−07	−24	14	05	32	−12	−42	63
25. Reflectiveness	−73	−76	−08	−12	13	03	−17	07
26. Scientism	−68	−58	−20	−24	13	04	−11	11
27. Sensuality	−55	−48	−31	−28	−05	05	−11	−01
28. Sexuality	07	−02	34	22	−09	−16	−45	73
29. Succorance	−51	−29	−43	−47	06	−17	−08	07
30. Understanding	−60	−56	−30	−28	25	07	−15	02

3. *Compulsivity and restraint*—a tendency to be cautious, orderly, and reflective in conforming to social and intellectual demands. (Factor III)

4. *Heterosexual dominance*—a tendency to desire the athletic, social, and intellectual qualities for heterosexual dominance or popularity. (Factor IV)

REFERENCES

Kaiser, Henry F. The Application of Electronic Computers to Factor Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 141–151.

Stern, George G., Stein, Morris I., and Bloom, Benjamin S. *Methods in Personality Assessment.* Glencoe, Illinois: The Free Press, 1956.

Wolf, Richard. The Measurement of Environments. *Proceedings of the 1964 Conference on Testing.* Educational Testing Service, 1965.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

PREDICTING GRADUATE SUCCESS AT WINONA STATE COLLEGE

CONSTANCE M. ECKHOFF
Winona State College

Problem

The purpose of this study was to investigate the relationship between selected background variables and achievement in graduate courses at Winona State College and to determine how accurately these variables predict achievement.

Background Variables

The background variables selected for investigation were (1) undergraduate grade point average (UGPA), (2) Miller Analogies Test scores (MAT), and (3) Advanced Education Section (AE) scores from the Graduate Record Examinations.

Criterion

The criterion was overall graduate grade point average (GGPA) in which grade points were assigned to marks as follows: A = 3, B = 2, C = 1, D = 0, and E or F = −1.

Statistical Methods

Stepwise multiple regression analysis was used to investigate the relationship among the background variables and the criterion. With this method, the matrix of correlations among predictor variables was inverted, and beta weights for each of these variables were calculated. A '*t*' test was used to test the significance of each beta weight against zero at the .05 level. The predictor variable that yielded the beta weight with the smallest '*t*' value was deleted from

the inverse of the matrix at each step until all weights reached significance. Beta weights were then calculated for the remaining predictor variables through use of the new inverse.

Subjects

The subjects were 185 secondary education majors and 111 elementary education majors with thirty or more quarter hours accumulated at Winona State College.

Results

Tables 1 and 2 present the results of stepwise multiple regression analysis for secondary and elementary majors, respectively. Two variables, UGPA and MAT, yielded significant regression weights for predicting GGPA for secondary majors. This optimum regression function yielded a multiple correlation coefficient of .51. Thus, although the content of the Advanced Education Section does appear relevant to graduate work in education, the test did not have a high degree of relationship with graduate achievement. When this variable was deleted after the first step of analysis, no great decrease in the multiple correlation coefficient resulted. For elementary majors, two variables, UGPA and AE, yielded significant regression weights for predicting GGPA. This optimum regression function yielded a multiple correlation coefficient of .30. The MAT, which did not yield a significant degree of relationship with graduate success for these subjects, was deleted after the first step of analysis. No decrease in the multiple correlation coefficient occurred.

Conclusions

This investigation of the relationship of selected background variables to graduate grade point average suggested the following conclusions:

1. Optimum prediction of graduate success for secondary education majors can be made by using a least-squares regression function containing background variables of undergraduate grade point average and a score on the Miller Analogies Test.

2. Optimum prediction of graduate success for elementary education majors can be made by using a least-squares regression function containing background variables of undergraduate grade point av-

erage and a score on the Advanced Education Section of the Graduate Record Examinations.

It is recognized that additional cross-validation studies should be undertaken with new groups of students. Such investigations have been planned for future classes of students.

TABLE 1

Beta Weights and Multiple Correlation Coefficients Before and After Deletion of Non-Significant Variables for Secondary Majors. (N = 185)

	Variable			
Statistic	AE	UGPA	MAT	$R_{1.23\ldots j}$
Beta	.1133	.4053	.1639	.52
Beta		.4070*	.2335*	.51

* $p < .01$

TABLE 2

Beta Weights and Multiple Correlation Coefficients Before and After Deletion of Non-Significant Variables for Elementary Majors. (N = 111)

	Variable			
Statistic	MAT	UGPA	AE	$R_{1.23\ldots j}$
Beta	−.0007	.2142	.1811	.30
Beta		.2141*	.1807**	.30

* $p < .01$
** $p < .05$

REFERENCE

Eckhoff, Constance M. Predicting Graduate Success at Winona State College. Unpublished Master's thesis, Winona State College, Winona, Minnesota, 1965.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

PREDICTING ACADEMIC PERFORMANCE IN A SMALL SOUTHERN COLLEGE

M. K. DISTEFANO, JR.[1]
Central Louisiana State Hospital
Pineville, Louisiana

AND

MARY L. RICE[2]
Louisiana College

Purpose

The purpose of this research was to study the validity of the School and College Ability Test (SCAT) in predicting academic performance at Louisiana College, Pineville, Louisiana.

Predictor

The SCAT (Form 1A) was the predictor variable studied. Constructed by Educational Testing Service, the SCAT is an aptitude test designed to aid in estimating the first year college applicant's ability to perform college level work. The test, which measures school-related abilities of a verbal and quantitative nature, yields three scores: (1) Verbal, (2) Quantitative and (3) Total.

Criterion

Grade point average served as the criterion in the study. Both first-year grade point average and overall four-year grade point average were employed as measures of academic performance.

---

[1] Full-time staff psychologist at Central Louisiana State Hospital and part-time assistant professor of psychology at Louisiana College.

[2] Now at the University of North Carolina, Greensboro, North Carolina.

Subjects

The SCAT was administered to 698 entering college freshmen and covered a five year period.

Results and Discussion

Pearson product-moment correlations between first-year grade point average and the Verbal, Quantitative, and Total scale scores were .48, .16, and .48, respectively ($N = 698$). The Verbal and Total scales were found to be significantly correlated at the .01 level, but the correlation for the Quantitative scale was not statistically significant.

SCAT scores were correlated with each student's four-year grade point average with scores available on 110 students. Pearson product-moment correlations for the Verbal, Quantitative, and Total scales were .68, .38, and .61, respectively. Although the correlations were all statistically significant, the Verbal scale was the best predictor of four-year grade point average.

Several prior studies have disclosed significant positive correlations between all three SCAT scores and grade point average (Mann, 1961; Vick and Hornaday, 1962; Lewis, 1962). However, there have been conflicting results with the Quantitative scale. For example, Boyce (1964) reported that the SCAT Quantitative score was not a significant predictor of college mathematics grades. Endler and Steinberg (1962) found that all three SCAT scales were significantly correlated with the first year grade point average correlations for male students. The present study found the Verbal Scale of the SCAT to be a much better predictor of academic performance than the Quantitative scale.

REFERENCES

Boyce, Richard W. The Prediction of Achievement in College Algebra. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 419–420.

Endler, Norman S. and Steinberg, Danny. Prediction of Academic Achievement at the University Level. *Personnel and Guidance Journal*, 1962, 41, 694–699.

Lewis, John W. Comparing Zero-Order Correlation from SCAT Total and Multiple Correlation from SCAT Quantitative and Verbal at Southern Illinois University. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 397–398.

Mann, Sister M. Jacinta, S.C. The Prediction of Achievement in a Liberal Arts College. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 481–483.

Vick, Mary Catherine and Hornaday, John A. Predicting Grade-Point Average at a Small Southern College. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 795–799.

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

EFFECTS OF ANSWER-SHEET FORMAT ON ARITHMETIC TEST SCORES

HENRY F. DIZNEY
PHILIP R. MERRIFIELD
AND
O. L. DAVIS, JR.
Kent State University

Lindquist (1964) has suggested two general criteria against which the design of answer sheets for tests may be judged:

1. The design should minimize the effect of irrelevant factors on test scores, i.e., mechanics of answer sheet use must not influence the test's validity; and
2. Within the limits of the first criterion, the answer sheet should be as convenient to administer and mark, and as economical to use as possible.

Lindquist properly emphasized the first criterion. Its implications would naturally vary with the nature of the examinees and scoring system as well as with the purpose of the test. Despite the obvious importance of the suggested criteria, the widespread use of separate answer sheets, and various methods of scoring, little empirical evidence is available concerning the effects of answer sheet format upon test results or examinees. One of the few studies along these lines was completed recently by Miller (1965). He investigated the effects of altering both color and marking position of answer sheets upon Lorge-Thorndike Intelligence Test scores. He studied these effects at three grade levels in four cities and concluded that both color and answer format had non-significant effects upon test results for eighth and twelfth graders. For fourth graders, significant interaction effects prevented a neatly drawn interpretation. Clearly,

theoretical and practical questions persist about the effects on different kinds of tests of varying answer sheet formats.

The present study was conducted in contemplation of a change from IBM 805 to IBM 1230 scoring systems in a university testing center. The two scoring systems require different answer sheet formats, the 805 being generally well-known by most students. The 805 format has response blocks numbered in columns from top to bottom of the page; the 1230 format, new to most students at this time, has answer blocks numbered in rows from left to right.

Procedure

It was decided to compare the performance of students using each format on the same test. The test chosen was one routinely administered to all sophomores in the College of Education as an arithmetic proficiency test. Failure on the test necessitates enrolling in a non-credit remedial course designed to alleviate the deficiency. Using this test seemed to insure two desirable conditions: (1) an unselected group; and (2) one whose members were motivated to perform well. The problems in the test were arrayed in columns and numbered from top to bottom; the 805 format was, in a sense, favored by this arrangement of items.

A brief questionnaire was designed to elicit affective responses concerning the answer sheet used for the test. Copies of the two types of answer sheets were distributed at random along with the test booklets and questionnaires to 499 students presenting themselves for examination early in the fall quarter 1964. As a result of the exigencies of assigning students to testing groups, 240 responded to the 805-format answer sheet and 259 used the 1230-format answer sheet. It was believed that the assignment of answer sheets to students was essentially random with respect to all performance-related variables, and that the sample sizes were sufficiently large to mitigate alarm. Questionnaire responses and test scores were collated and analyzed.

Results

In response to each of three questions, proportionately more students using the 1230 format reported difficulty in using the answer sheet than did those students using the 805 format. However, a median test of the scores of those reporting difficulty using the two

formats indicated no significant differences. In Table 1 the chi-square values for each question are presented; values in column A refer to frequency of reported difficulty, and are all significant at the .001 level; values in column B resulted from the median test applied to scores, and are not significant.

TABLE 1

Summary of Statistical Analysis

Question	A[a]	B[b]
Difficulty in:	$\chi^2(df = 1)$	$\chi^2(df = 1)$
a. using the answer sheet	89.	1.65
b. orienting to the arrangement of answer blocks	142.	.01
c. coordinating problems and answer blocks	41.	2.90

[a] χ^2 value for frequency of users reporting difficulty in two format groups.
[b] χ^2 value, median test, for scores obtained by those reporting difficulty in two format groups.

Students using the 1230 format very definitely felt greater difficulty in the testing situation. Nevertheless, in terms of their test performances, under conditions of high motivation, format mattered not.

REFERENCES

1. Lindquist, E. F. Basic Considerations in Answer Sheet Design. *Testing Today:* Houghton Mifflin Company Circular, 1964.
2. Miller, Irwin. A Note on the Evaluation of a New Answer Form. *Journal of Applied Psychology,* 1965, 49, 199–201.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

SOCIO-ECONOMIC BACKGROUND AND FAILURE IN THE HIGH SCHOOL EXAMINATION

S. L. CHOPRA
Lucknow, India

ONE of the serious problems facing educators in India is the rather high percentage of failures in the high school examinations. In the year in which the present study was conducted 53.85 per cent of the high school examinees in Uttar Pradesh had failed. In the present study, therefore, an effort was made to see (Coster, 1959) how far these failures were related to socio-economic background and (Hohne, 1949) whether the differences in the percentages of failures in different socio-economic groups could be accounted for by variations in levels of measured intelligence.

Procedure

The test of Progressive Matrices was administered to 1,359 randomly selected high school students (age range 14–17) studying in twenty-two urban and six rural secondary schools in Lucknow district. Data for parental occupation and results in the high school examination were also collected.

Results

In Table 1 the entries show the percentages of failures in the different occupational groups. It is apparent that the differences between the number of failed students in the various occupational groups were statistically significant. Figures in column four show that there was a gradual rise in the percentage of failures as one moves from the higher to the lower occupational groups.

Were the higher percentages of failures in the lower occupational

TABLE 1

Percentages of Failures in the Different Occupational Groups

Parental Occupation	Number of Students	Failures No.	Failures %
(1)	(2)	(3)	(4)
1. Prof., Admn., Exec. & Managerial	186	50	27
2. Clerical	268	138	51
3. Minor Business & Sales Workers	226	124	55
4. Skilled Workers	240	133	55
5. Agriculturists	320	190	59
6. Unskilled Workers	119	72	61
Total	1,359	707	52

$X^2 = 28.48$ ($p < .01$)

groups due to lower intellectual level of students in these groups? To answer this question, six groups of 81 students each, from the six occupational groups, were matched in intelligence test scores (person for person). In Table 2 the percentages of failures in these

TABLE 2

Percentages of Failures in the Different Occupational Groups Matched for Intelligence

Parental Occupation	Number of Students	Failures No.	Failures %
(1)	(2)	(3)	(4)
1. Prof., Admn., Exec. & Managerial	81	19	23
2. Clerical	81	41	51
3. Minor Business & Sales Workers	81	42	52
4. Skilled Workers	81	46	57
5. Agriculturists	81	49	60
6. Unskilled Workers	81	49	60
Total	486	246	51

$X^2 = 15.57$ ($p < .01$)

six matched groups are shown. From the entries in Table 2 it appears that the differences in the numbers of failures in the various occupational groups were statistically significant even when the individuals in these groups were matched for intelligence test scores. Figures in column four show that when intellectual level was controlled there was also a gradual increase in the percentages of failures as one moves from higher to the lower occupational groups.

These results suggest that socio-economic background is positively related to success in the high school examination and that even when measured intelligence is held constant the richer the socio-economic background the less is the probability of failure.

REFERENCES

Coster, John K. Some Characteristics of High School Pupils from Three Income Groups. *Journal of Educational Psychology*, 1959, 50, 55–62.

Hohne, H. H. The Prediction of Academic Success. *Australian Journal of Psychology*, 1949, 1, 38–42.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

PREDICTION OF GRADES IN GRADUATE EDUCATION COURSES

L. L. AINSWORTH AND A. M. FOX
Sam Houston State College

Problem

THE purpose of this study was to determine the extent to which Miller Analogies Test (MAT) scores may be useful for the prediction of grade-point ratios (GPR's) in graduate Education courses.

Sample

The sample consisted of 1,413 students at Sam Houston State College for whom MAT scores and at least one grade in a graduate Education course were available, for the period July, 1959, through August, 1964.

Procedure

GPR's were calculated, and then relationships of MAT to GPR's were determined by the Stepwise Regression Analysis Program (STRAP) in an IBM 1620 Data Processing System. An F-ratio of 4.0 was used for the criterion in testing the significance of the relationships. Significant relationships were found for MAT scores with total GPR (GPRT), basic required courses (GPRB), Principles of Guidance (GPR33), Measurement and Evaluation (GPR93), Secondary Curriculum (GPR94), and Human Growth and Development (GPR97). Regression equations, coefficients of alienation (k), and indices of forecasting efficiency (E) were then computed for these significant relationships.

Results

The results of this study are summarized in Table 1.

TABLE 1

Relationships of MAT to Dependent Variables

Variable	N	r	k	MAT Mean	MAT S.D.	GPR Mean	GPR S.D.	Regression Equation	Standard Error of Estimate σyx	E
GPRT	1,413	0.31	0.95	32.5	13.4	3.1	0.56	2.7 + 0.013(MAT)	0.54	5
GPRB	1,185	0.30	0.95	32.6	13.2	3.0	0.64	2.5 + 0.015(MAT)	0.61	5
GPR33	349	0.28	0.96	33.8	13.6	3.1	0.76	2.6 + 0.016(MAT)	0.73	4
GPR93	957	0.31	0.95	32.4	13.2	3.0	0.71	2.4 + 0.017(MAT)	0.67	5
GPR94	539	0.19	0.98	33.6	13.6	3.1	0.66	2.8 + 0.009(MAT)	0.65	2
GPR97	640	0.25	0.97	32.8	13.3	3.0	0.68	2.6 + 0.013(MAT)	0.67	3

Educational and Psychological Measurement
Vol. XXVI, No. 2, 1966

PREDICTIVE RELATIONSHIPS BETWEEN ITEMS ON THE REVISED STANFORD-BINET INTELLIGENCE SCALE (SBIS), FORM L-M, AND TOTAL SCORES ON RAVEN'S PROGRESSIVE MATRICES (PM), BETWEEN ITEMS ON THE PM AND TOTAL SCORES ON THE SBIS, AND BETWEEN SELECTED ITEMS ON THE TWO SCALES[1]

E. GEORGE SITKEI
Pacific State Hospital
AND
WILLIAM B. MICHAEL
University of California, Santa Barbara

Although Raven's Progressive Matrices (PM) test was originally developed in England, it has been widely used in America for a number of years. Relatively little systematic research has been conducted concerning its relationship to well known scales of intelligence developed in the United States. In his review of literature pertinent to the PM, Burke (1958) noted that only a few studies had been concerned with the validity of the scale or with its item characteristics. Two other studies which reported correlations between scores on the PM and scores on parts of other intelligence tests were those of McLeod and Rubin (1962) and Martin and Wiechers (1954).

Purpose

It was the purpose of the present study to report for separate samples of 63 men and 80 women as well as for the total sample

[1] Supported in part by the National Institute of Mental Health Grant No. MH08667: Socio-Behavioral Study Center for Mental Retardation, Pacific State Hospital, Pomona, California. Computing assistance was obtained from the Health Sciences Computing Facility, UCLA, sponsored by NIH Grant FR-3.

of 143 adults the following information: (1) a brief summary of the range and degree of the biserial coefficients of correlation of items at or above the eighth year level in the Revised Stanford-Binet Intelligence Scale (SBIS), Form L-M, with total scores in the PM, (2) a comparable summary of the biserial coefficients found between each item in the PM with total scores on the SBIS, and (3) a short interpretative statement concerning any pattern of relationships associated with the 25 highest intercorrelations (phi coefficients) found between any one of the items on the SBIS and any one of the items on the PM. The biserial coefficients of correlation were calculated on an IBM 7090 computer through use of the IBS-003 biserial and point biserial program at the Western Data Processing Center at UCLA. The phi coefficients were derived from the BIMED 02D program.

Sample

The total sample studied consisted of individuals ranging in age from 16 to 49 who were tested in their homes as participants in a community-wide survey concerned with identification of mentally retarded individuals in a city of about 100,000 population in Southern California (Mercer, Dingman, and Tarjan, 1964). Of a pre-selected 10 per cent stratified sample of a total group of home dwellers in the community, interviewers had previously visited successfully 2,661 of the pre-selected households, a number which represented about 91 per cent of the pre-selected group. During and subsequent to his visit in a household, the interviewer completed a questionnaire for each person and obtained information on his socio-developmental history. In order that a substantial range of ability level might be realized, two random samples were obtained: (1) one of persons who scored in the lowest 10 per cent on the socio-developmental scale, and (2) one of all persons in the interviewed households.

These two samples were combined in such a manner as to obtain a distribution of scores in the two ability measures so that a proportional stratified sample would result. As defined by the selection procedures, based on the socio-developmental scale employed in the community survey, a distribution of frequencies would probably be approximately rectangular relative to the established percentile norms for the scale. Thus, with respect to an

existing normative C-scale of 10 steps for the socio-developmental measure the number of individuals selected for testing at high, medium, and low portions of this measure—steps 8, 9, and 10 for high; steps 4, 5, 6, and 7 for medium; and steps 1, 2, and 3 for low—were 59, 89, and 64, respectively. Actually tested at each of these levels were 50, 70, and 48 individuals, respectively. In relation to intelligence, the actual distribution for the sample appeared to be more heterogeneous than for the general population as evidenced by the fact that the standard deviation for the SBIS was 24.81 instead of the familiar value of 16.00 found on the national standardization.

The composition of the sample from the standpoint of sex, age, and ability level is described in Table 1. For each sex group as well as for the total sample, descriptive statistics regarding chronological age, IQ scores on the SBIS, scores on the PM, and intercorrelations of the total scores on the SBIS and PM are furnished.

TABLE 1

Descriptive Statistics Regarding Chronological Age, IQ Scores on the Revised Stanford-Binet Intelligence Scale (SBIS), Form L-M, Scores on Raven's Progressive Matrices (PM), and Intercorrelations of the Two Tests for the Two Samples of Men and Women and for the Total Sample

Sample	N	Chronological Age Data		SBIS IQ		Raven's PM Scores: Raw Scores		Raven's PM Scores: Centile for Mean	Correlations of Two Scores
		M	*Range*	*M*	σ	*M*[a]	σ		*r*
Men	63	35.1	16–49	110.2	23.2	37.5	12.2	42.5	.67*
Women	80	33.2	17–49	103.7	23.7	35.6	12.3	37.5	.62*
Total	143	34.1	16–49	106.0	24.8	36.0	12.3	40.0	.65*

* Significant beyond .001 level.
[a] The means constitute interpolated values from percentile data available in the test manual.

Findings

From the coefficients of correlation between the SBIS and the PM, which for the male, female, in total samples are seen in Table 1 to be .67, .62, .65, respectively, it is apparent that the two scales probably contain a considerable amount of common-factor variance. Further support for such a conclusion is also evident from the summary of item analysis data in Table 2, in which the frequency distributions of the biserial coefficients of correlation for the SBIS items against total scores on the PM and for the PM

items against total scores on the SBIS are reported. (Although not presented because of space limitations, the difficulty level of each item in the two scales as well as the coefficient of biserial correlation of each item against total score on the scale of which it is not a contributor are available.)[2] It is evident that the degree of asso-

TABLE 2

Frequency Distributions of Biserial Coefficients of Correlation Obtained in Item Analyses

Size of Biserial Coefficients	SBIS Items with Total Scores on PM			PM Items with Total Scores on SBIS		
	Men ($N = 63$) f	Women ($N = 80$) f	Total Sample ($N = 143$) f	Men ($N = 63$) f	Women ($N = 90$) f	Total Sample ($N = 143$) f
.80 or higher	1			1		
.75–.79	1	1		2		2
.70–.74	7	0	1	2	8	3
.65–.69	12	6	9	4	2	1
.60–.64	11	15	16	4	1	3
.55–.59	12	19	17	7	4	11
.50–.54	13	6	11	5	13	8
.45–.49	3	6	3	8	4	4
.40–.44	0	5	7	8	8	11
.35–.39	4	2	2	2	5	7
.30–.34	3	3	2	2	2	3
.25–.29	1	1		4	4	2
.20–.24		1		1	1	0
.15–.19		0		0	3*	1*
.10–.14		0		1*	0	1*
.05–.09		0		0	0	3*
.00–.04		3*		9*	5*	
	68	68	68	60	60	60

* Small size of the coefficients is attributable to the fact that the proportion passing the item was 1.00 or approximately 1.00 or that the proportion was 0.00 or close to 0.00.

---

[2] A limited number of copies of the following information is available on a loan basis from the senior author: (1) a table (Table 3) that provides a code for identification of the 68 items of the SBIS starting with the eighth year level and terminating with the third level of the superior adult as well as a code for the PM test with its five blocks or sets of items A, B, C, D, and E (12 items per block); (2) three tables (Tables 4.1, 4.2, and 4.3)—one for the male sample, one for the female sample, and one for the total sample—furnishing the difficulty level of each item at or above the eighth year level in the SBIS, Form L-M, as well as the biserial coefficient of correlation of each item with total scores on the PM; (3) three similar tables (Tables 5.1, 5.2, and 5.3) presenting the difficulty level of each item in the PM as well as biserial coefficient of correlation of each of these items with total scores on the SBIS; and (4) a table (Table 6) reporting the 25 highest intercorrelations (phi coefficients) found between 25 coded items in the PM and 25 coded and briefly described items in the SBIS.

ciation of items in one intelligence scale with the total scores earned by the examinees in the other scale was relatively high for each of the three groups studied. That a small number of items showed close to zero correlation with total score in the scale of which these items were not members was probably a statistical artifact arising from proportions either of unity (or near unity) or of zero (or nearly zero) answering the item correctly. Thus for very easy or for very difficult items the coefficients were markedly attenuated.

When a search was made for patterns or combinations of items in one scale that registered high correlation with the total scores of the other scale, it was difficult to determine any clear-cut clusters of items that were associated with either a given range of correlation coefficients or with sex membership. In a detailed study of complete item-analyses which are available elsewhere (see footnote 2) there was a slight suggestion from the magnitudes of the biserial coefficients for certain items on the SBIS with total scores on the PM that the males tended to use verbally oriented abstract reasoning in solving items on the PM, whereas the women seemed to rely somewhat less on verbalization and somewhat more on symbolic induction than did the men. Examination of the coefficients of PM items with total scores on the SBIS revealed no consistent pattern of magnitude from which any clear-cut implications could be drawn.

Clearly, one of the most promising avenues for formulation of inferences regarding the psychological processes underlying responses to the PM items would be a factor analysis of the intercorrelations of PM and SBIS items or of short subtests of items taken from these scales—subtests of sufficient reliability that could be arranged or contrived to test certain hypotheses of cognitive function. Since a detailed examination of the 25 highest intercorrelations (phi coefficients) of items in the SBIS with items in the PM also did not reveal any meaningful clusters that could be given a psychological interpretation, the need for factor analytic studies is indeed quite apparent.

Summary

For two adult samples of 63 males and 80 females and for the combined sample of 143 adults, a study of biserial coefficients of correlation of items in the Revised Stanford-Binet Intelligence Scale

(SBIS), Form L-M, with total scores on Raven's Progressive Matrices (PM) scale as well as of the items in the PM with total scores on the SBIS revealed a relatively high degree of communality for the two instruments—a fact which was also supported by correlation coefficients of .67, .62, and .65, respectively, between total scores on the two instruments for the male, female, and total samples. The failure of an examination of the item-analysis data and of the 25 highest intercorrelations of items from the two scales to show any clear-cut clusters amenable to psychological interpretations pointed to the need for factor analyses of the intercorrelation of items or of subtests of items from the SBIS on PM scales.

REFERENCES

Burke, H. R. Raven's Progressive Matrices: A Review and Critical Evaluation. *Journal of Genetic Psychology*, 1958, 93, 199–228.

Martin, Anthony W. and Wiechers, James E. Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology*, 1954, 18, 10.

McLeod, Hugh N. and Rubin, Joseph. Correlation between Raven Progressive Matrices and the WAIS. *Journal of Consulting Psychology*, 1962, 26, 190–191.

Mercer, Jane R., Dingman, Harvey F., and Tarjan, George. Involvement, Feedback, and Mutuality: Principles for Conducting Mental Health Research in the Community. *The American Journal of Psychiatry*, 1964, 121, 3.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XXVI, No. 2, 1966

The Sixth Mental Measurements Yearbook by Oscar Krisen Buros, Editor. New Jersey: The Gryphon Press, 1965. Pp. xxxvi + 1714. $32.50.

Expanded considerably in size and quality in harmony with the objectives and purposes of previous *Yearbooks, The Sixth Mental Measurements Yearbook* not only includes the major organizational divisions of *The Fifth Yearbook* but also contains, contrary to the announced future policy in *The Fifth Yearbook,* comprehensive bibliographies about the construction, validation, and uses of specific tests and scales as well as both a detailed listing and reviews of tests in English-speaking countries outside the United States. In addition, a substantial and exceedingly expensive improvement is the presentation of information regarding the status of tests listed in *Tests in Print*, but not otherwise considered in *The Sixth Yearbook.* Thus for the first time, all 1219 tests known to be in print as of mid-1964 are cited; 795 critical test reviews by 396 reviewers are included along with 97 excerpts from reviews of tests that initially appeared in 30 journals; and 8001 references for specific tests are incorporated.

Information regarding the expansion of *The Sixth Yearbook* relative to *The Fifth Yearbook* may be noted as follows: (1) the number of tests cited, 1219, constitutes 27.4 per cent more entries; (2) the number of pages is 32.6 per cent greater; (3) 628 new tests are incorporated; and (4) an increase of 8.6 per cent exists in the number of books listed in the section "Books and Reviews." Along with a modest rise in price the total number of pages has jumped from 1292 in *The Fifth Yearbook* to 1714 in *The Sixth Yearbook* (excluding 36 introductory pages).

Another interesting set of statistics is that of the frequency of tests by major classification. The breakdown is shown on the following page.

Despite the small percentage of multi-aptitude and achievement batteries in the total number of tests reviewed, the annual volume of such tests administered throughout the United States may well exceed that found for any of the other categories. Because of this comprehensive use Buros assigned a great deal of space to the review coverage of these two types of batteries.

Classification	Number	Percentage
Personality	196	16.1
Vocations	179	14.7
Intelligence	131	10.7
Miscellaneous	108	8.9
English	99	8.1
Mathematics	96	7.9
Reading	87	7.1
Foreign Languages	77	6.3
Science	74	6.1
Social Studies	61	5.0
Business Education	29	2.4
Achievement Batteries	28	2.3
Sensory-Motor	27	2.2
Multi-Aptitude	16	1.3
Fine Arts	11	0.9
Totals	1219	100.0

One additional set of statistics is that presented by Buros relative to the percentages of tests that are new or revised. As mentioned previously 628 tests were new—a number which exceeded the 445 (or 36.5 per cent) revised since the manuscript stage of *The Fifth Yearbook*. Thus the total number of new, revised, or supplemented tests was 1073 (or 88.0 per cent).

In the spirit of previous *Yearbooks* the test reviews are for the most part critical but constructive in their presentation. As one would reasonably expect, inadequacies in the manuals relative to information furnished about validity, reliability, and standardization are forcefully set forth again and again. In addition to the valuable sections on test reviews and on books and reviews, other helpful portions of *The Sixth Yearbook* include a periodical directory and index; a publishers directory and index; an index of test titles; an index of names of test authors, reviewers, and authors of articles about tests; and a classified index of tests—both those reviewed in this volume and those known to be in print.

An editorial masterpiece and a beautifully printed volume which is a tribute to the assiduous efforts expended by Oscar Buros over a period of nearly six years, *The Sixth Mental Measurements Yearbook* is, as one would expect from the Buros tradition, the best one to date. A sampling of test reviews from *The Sixth Yearbook* suggests that the critics are no less demanding than they have been in previous years, but probably are somewhat more aware of advances in statistical methodology as well as somewhat more alert to problems of construct validation, especially in relation to personality tests. There is little doubt that the person who reads conscientiously the reviews of any test that he may be considering for possible use will exercise much greater care than he would have exhibited otherwise had he not been exposed to the generally competent and detailed evaluations of the test reviewer.

As on previous occasions, Oscar Buros has expressed disappointment with the impact which the *Yearbooks* have had upon the development of discriminating taste on the part of consumers of tests. The message which *The Sixth Yearbook* carries apparently must be conveyed and supplemented by people who are in a position to communicate knowledge about basic standards for educational and psychological tests to their consumers or probable consumers. Competent individuals in the public schools who hold positions of supervisory and administrative responsibility may be expected to become increasingly more knowledgeable about standards of reliability, validity, norms, and directions for administering tests—especially as educational training in leading colleges and universities is improved. But perhaps the group of individuals who have the greatest responsibility are professors of psychology and education who in their classes and inservice workshops conducted off campus have the almost missionary obligation to reach as many potential consumers of tests as they possibly can. Once students have gained the necessary basic knowledge about the requirements of good tests they should be directed to read and to reread the Buros *Yearbooks* prior to making any decisions regarding the adoption of standardized tests for use in schools, industry, or clinical settings. Perhaps in a few years, with such concerted efforts on the part of professional leaders in educational and psychological testing, Oscar Buros' projected *Seventh Mental Measurements Yearbook* will accomplish for the rank and file test user what he so ardently had hoped would be achieved many years ago with earlier editions of the *Yearbooks*.

JOAN J. BJELKE
University of Southern California
AND
WILLIAM B. MICHAEL
University of California, Santa Barbara

Porteus Maze Test: Fifty Years' Application by Stanley D. Porteus. Palo Alto, California: Pacific Books, 1965. Pp. 320.

Stanley Porteus' life by almost any standard has been a highly successful one. He was vouchsafed fairly early an original insight in psychometrics and has carefully nursed his brain child along for over fifty years of application, research, and controversy. He developed his maze test in 1913, described its first application at the British Association for the Advancement of Science meeting at the University of Melbourne in August, 1914, and published his first papers on the topic the following year. The inception of the Porteus Maze Test and World War I was coetaneous.

In 1919 Porteus came from Australia to succeed Goddard at the Vineland Institute in New Jersey; then a few years later he went to the University of Hawaii, from which he has long retired as pro-

fessor emeritus. He has never been accorded adequate professional recognition among American psychologists, possibly because he was something of an Ausländer among the natives. It seems more probable, though, that his espousal of tests as a touchstone for evaluating racial differences in intelligence in the long dead controversies of a generation back had more to do with the lack of acceptance of his mazes and the claims he has made for them. One should think that a racial melting pot such as Hawaii would find such a set of attitudes especially difficult to empathize into or accept. Finally, the mazes ran head on into the Binet-type test which had only recently stirred up excitement in the United States prior to Porteus' arrival at Vineland. For 20 years after its publication in 1916 the Stanford-Binet Intelligence Scale was accepted as the true measure of intelligence, against which any newcomer must be validated. Thus it can be seen that Porteus and his mazes gave many hostages to fortune while making a bid for professional recognition in the United States: Porteus was "foreign," he chose unpopular causes, and Terman's Stanford-Binet Intelligence Scale was already entrenched. Only through long life, energy, and persistence have these hostages been neutralized. This book tells the story—a successful one—of the tangled Odyssey of the mazes.

The mazes began, as did the 1905 Binet scale, with a practical problem, the screening out of feebleminded children so that (it was hoped) they could more effectively be instructed in specialized schools. Porteus felt in 1913 that Goddard's version of the Binet Scale was a very imperfect measure: the mazes were developed to aid in the differential diagnosis of the feebleminded. Porteus believed that the performance measure he invented in the maze test cut through the pitfalls of the purely verbal measure and got at some native aptitudes of a practical nature, especially of "planfullness." It has taken him a lifetime to document his original hunch, but he has done it in a number of specific ways.

Aside from Porteus' own publications, one of the good early confirmations of the special extra quality to be found in the mazes but not in the Binet type test was a study of Poull and Montgomery in 1929. They gave the Stanford-Binet Intelligence Scale and the mazes to inmates of a youth facility on Randall's Island, New York, about one half of whom consisted of juvenile delinquents while the other half was relatively free of psychopathic tainting. Stanford-Binet Intelligence Scale showed no mean differences between the two groups; the mazes did—the juvenile delinquent group scoring nine points lower. Slowly here, more readily in England, the mazes began to be accepted as an adjunct test in the clinical armamentarium and increasingly often as the test of choice if the child or adult were non-English or nearly so.

About 25 years ago Porteus devised a qualitative way of scoring

the mazes which he published—a scoring system showing marked validity as a means of discriminating prison populations from more nearly normal groups. There has been repeated confirmation of Porteus' original findings. One of the more noteworthy was published in 1954 by Docter and Winder from Stanford. The mazes would appear to measure something both quantitatively and qualitatively which discriminates against the impulsive individuals so frequently found among the delinquent and criminal.

The Porteus Maze Test was found by Porteus and Peters, later by Landis and Zubin to be a sensitive measure of the cerebral insult of psychosurgery. Whereas the Wechsler-Bellevue Intelligence Scale showed no impairment subsequent to the operation, the mazes did. And Porteus has subsequently made the chilling point that if the practice effect on his mazes is taken into account, the damage is permanent in most but not in all cases. It is fortunate that this pioneering blind alley of the neurosurgeons was so soon tossed into the dustbin of history through the advent of chemotherapy.

Porteus also demonstrated from the beginning that the maze test is a measure much to be preferred to the traditional intelligence test when institutional groups—the feeble of mind, the criminal, or the delinquent—are to be sorted for training or treatment. It seems likely that the planfullness aspect, presumably calibrated by the mazes, is the significant parameter in its evident superiority as a test to the Binet and the Wechsler type of measure in these institutional areas.

In several anthropological expeditions, some by Porteus, some by others, the mazes have been found to be a type of measure which pre-literate groups accept and find interesting. Validation of the mazes for such groups is, of course, fragmentary. What has been reported makes it seem likely that the test is a valuable adjunct for the anthropologist to know about and to employ for research purposes in his investigations.

Most recently Porteus has devised a conformity-variability (C-V) for his mazes, a projective kind of measure which quantifies the degree to which the same kind of attack is used on the mazes when they are re-taken. Individuals can be matched successfully, he says, in about 90 per cent of the cases. If the conformity score is too high, Porteus feels that the person is constricted, meticulous, and compulsive; too low, and he suspects a varying amount of disorganization. Here are some hypotheses by the youthful octagenarian—those in need of a problem, let them try their hand at testing these brain children of the Ancient of Days from Aloha Land.

In 99 out of 105 studies Porteus has found the male to have a maze test mean superior to that of the female. Porteus does not exactly claim that this shows general male superiority of intelligence; he suggests that it probably represents inherent tempera-

mental differences. Since he has so often equated temperamental differences with manifest intelligence, one wonders whether he does not feel about women's mental inferiority somewhat the way he did about the various non-Caucasian groups in Hawaii and their numerous crosses which he tested and found wanting a generation ago. One hastens to add that this claim is not made in this latest book of Porteus. It may be that he has succumbed to the Zeitgeist. It may also be that he feels it is better not to tilt at windmills without allies.

We should be thankful to Porteus for his mazes and for his research. He is still a vigorous, undaunted sample of what originality, persistence, hard work, and long life can bring about. Let us hope that he brings us up-to-date once more when he reaches the century mark.

William D. Altus
University of California, Santa Barbara

Statistical Inference I, by Jerome C. R. Li. Ann Arbor Michigan: Edwards Brothers Inc. 1964. Pp. xix + 658.

Statistical Inference appears in two long volumes. Only *Statistical Inference I,* the revised edition of *Introduction to Statistical Inference,* which was published in 1957, is reviewed here.

Principal changes in the revised edition include a very welcomed improvement in printing and the unexplained loss of excellent sets of questions for most of the chapters. (Sets of exercises with solutions have been continued.) Two new chapters appear at the end of the new edition. A third review chapter has been added and is followed by a chapter which discusses outlying observations, homogeneity of variances, and determination of sample size.

The book is a non-mathematical exposition of the theory of statistics written for experimental scientists. Many of the examples in the text and in the exercises are drawn from psychology and education. No more mathematics than a background in high-school algebra is required.

Throughout the book, major emphasis is placed upon a series of sampling experiments, employing up to 1,000 replications, which are used to verify statistical theorems. By comparing the results of the sampling experiments, which were conducted by Li's students, with the mathematical properties of statistical variables, physical meaning is given to many concepts.

A table of 5,000 random normal numbers with a mean of 50 and a variance of 100 is provided for the instructor who wishes to have his class conduct such sampling experiments. This valuable inductive approach deserves more attention and lends itself quite readily to the use of high-speed computers for increasing the number of replications in order to generate better empirical approximations.

Two review chapters divide the book into three parts. The first part (Chapters 1–13) of the book presents the basic concepts of statistical inference and introduces the normal, χ^2, t, and F distributions. Sections are devoted to the "u-tests" ("z-tests"), the "t-test" of paired observations, the "t-test" of independent groups, the χ^2 test of population variance, the "F-test" of two population variances, one-way analysis of variance, and confidence intervals.

This review is written after the reviewer's having used Li's book in several classes of intermediate statistics taken primarily by psychology majors. Students have been almost unanimous in their praise of the clarity of presentation of the material in part one. This remarkable state of affairs appears to be the result of the great care that Li has taken with sentence structure and transition and his decision to exclude many supporting and tangential topics in order to focus attention on the concepts at hand. One does not find any general introduction to probability, permutations and combinations, the rules of summation notation, and the topic of expected values. However, for the student who has just stumbled through a mathematically elegant introductory course in statistics, this book offers an opportunity to discover what statistics is designed to accomplish. The instructor in an introductory class will probably want to supplement and provide additional background material at some points.

One perplexing change of level of presentation is found in part one. In a well written chapter, the χ^2 distribution (chapter 7) is presented as a general model which is generated by repeatedly summing k independent values of u^2, where u is defined as a standardized normal deviate. The statistic SS/σ^2 is then mathematically shown to follow this general model. Once the necessary background has been presented, one would then expect that the t and F distributions would also be defined in similar terms and that the statistics that bear their names would be shown to conform to the general model. Instead, Li returns to the use of sampling experiments to convince the reader that the statistics have been properly named.

The second part (Chapters 14–20) presents the topics of randomized blocks, tests of specific hypotheses in the analysis of variance, linear regression, factorial experiments, and analysis of covariance. Repeated measurements designs are neither presented nor mentioned either with regard to three or more treatment conditions in this part of the book or in connection with the "*test-test*" of paired observations in part one.

The presentation of each topic includes a limited discussion of experimental design. This is a common practice in many useful statistics texts. However, a general discussion of experimental designs, perhaps as an overview before specific methods of analysis are presented, would help the student to develop a framework for the

methods of analysis. This would be particularly important for commonly used designs that otherwise are not covered.

Sampling experiments are used extensively to verify the sampling distributions of the statistics that are considered. The importance of an individual degree of freedom is stressed and helps to provide continuity from topic to topic. A basic part of Li's approach is to relate the separate topics. This is accomplished with considerable skill in two valuable sections, one which compares and contrasts linear regression and analysis of variance, and another section which discusses the relationships between analysis of covariance and factorial experiments.

The chapter on factorial experiments quickly represents the two dimensional case and discusses the differences between the fixed, mixed, and random models. An important error of omission occurs in that no discussion is offered on how to proceed in the fixed model case when the interaction is significant.

Part three (chapters 21–25) is primarily concerned with extending, through the vehicle of analysis of variance, most of the methods of analysis that were developed in the first two parts to sampling from one or more binomial populations. The chi-square test of goodness of fit for one or more samples from binomial populations and the chi-square test of independence for two binomial samples are shown to be directly related to analysis of variance with appropriate modifications and restrictions. Sampling experiments are used to show that the statistics that are computed from the chi-square test of independence for two or more multinomial populations and the chi-square test of goodness of fit for one multinomial populations approximately follow the χ^2 distribution. By using this approach, Li gives these topics comparatively thorough treatment in relatively short space and provides a link between sampling from normal and from binomial populations. The reader who has digested the topics of analysis of variance, linear regression, and the use of a single degree of freedom, can appreciate the degree of unity that Li has brought to the parametric and nonparametric methods that are discussed.

In summary, Li's book is a solid and eminently readable presentation of many topics that are often presented in introductory and intermediate courses in statistics. On the strength of the clarity or the exposition alone, the book can serve many functions. The instructor who covers the topics that are found in Li's book at the same mathematical level of presentation will find many excellent reasons for choosing this book. The student who wants to review the basic concepts and methods that are presented in the first part will find the presentation exceptionally lucid. *Statistical Inference I* is the logical choice for providing the necessary background for *Statistical Inference II* which deals with many difficult to find topics

on multiple regression and its ramifications. Finally, for the instructor who will probably never be completely satisfied with any one statistical text, Li's book can serve very satisfactorily in giving support to and in freeing classroom time for individual approaches and emphases in statistics.

EMANUEL LASK
California State College at Los Angeles

Exercises to Accompany Fundamental Statistics in Psychology and Education (4th Edition) by J. P. Guilford, William B. Michael, and Stephen W. Brown. New York: McGraw-Hill, 1965. Pp. vii + 204. $3.50.

This set of exercises and problems is essentially a collation of those from the third editions of the authors' *Elementary Statistical Exercises* and *Intermediate Statistical Exercises*. Some appropriate new material has been added with the intention of providing a single comprehensive accompaniment to the fourth edition of Guilford's *Fundamental Statistics in Psychology and Education.*

The aim of a good workbook should be to include as many relevant exercises and problems as possible, not to provide large numbers of beautifully laid out but uninformative worksheets. In this, Guilford, Michael, and Brown have succeeded very well. Instructions and directions for problem lay-out are clear but often implicit. The accounting approach to statistical computations is largely deemphasized in favor of a "what's going on?" approach. The exceptions to this are the inordinate attention paid to coding of observations and to the trial balance method of computing the mean and standard deviation, and the inclusion of large and clumsy (even if traditional) tables for calculating a product-moment correlation coefficient. With the availability of modern computing techniques and devices, these emphases cannot be justified. The section on analysis of variance is much less heavily weighted than the current interest and importance of this topic demand in a general introduction. However, the faults mentioned are more reflections of weaknesses in Guilford's fourth edition than of the workbook itself. Taken alone, *Exercises to Accompany* is an excellent workbook which should provide students very adequate operational experience in statistical procedures and statistical thinking.

JAMES A. WALSH
Iowa State University

Measurement and Evaluation in Teaching by Norman E. Gronlund. New York: The Macmillan Company, 1965. Pp. xii + 420. $6.95.

Gronlund's purpose is to introduce the reader

> . . . to the principles and procedures of evaluation which are essential to good teaching. The main theme which runs throughout the book is that evaluation is an integral part of the teaching-learning process and that it involves three fundamental steps: (1) identifying and defining instructional objectives in behavioral terms, (2) constructing or selecting evaluation instruments which most effectively appraise these specific learning outcomes, and (3) using the results to improve learning (p. vii).

Centering his book upon the place of evaluation in teaching, rather than upon principles of making and interpreting tests as though they were ends in themselves, Gronlund avoids the shortcoming of writers of introductory texts who leave to the immature student the task of integrating the concepts of the psychology of learning, methods of teaching, and tests and measurements. Gronlund is second to none in achieving this needed synthesis and succeeds in portraying evaluation *as an integral part of teaching.*

Emphasis upon computational work is avoided. Abilities in interpreting and applying statistical concepts that are essential to effective use of tests are the outcomes sought. Rank-order correlations and expectancy tables are introduced logically in connection with the discussion of validity. The normal curve and standard deviations are capably presented within the context of standard score norms; however, "since the method of computing the standard deviation is not especially helpful in understanding it, the procedure will not be presented here" (p. 286).

No attention is given to the history of measurement. The names of such pioneers as Binet, Terman, and E. L. Thorndike do not appear. While historical content is often interesting (to instructors, at least), its omission from a text which is properly concerned with improving the ability of teachers to evaluate their pupils seems well advised.

Part I deals with the evaluation process. Superb examples are given of developing behaviorally stated objectives for a variety of subject fields that range from primary to high school levels. Consistent with the view that course "content serves its most useful purpose when viewed as a means of obtaining educational objectives rather than as an end in itself" (p. 21), the second chapter develops skills in defining educational objectives clearly and behaviorally. This chapter, which is probably unexcelled in teaching the skills of writing objectives, should warm the hearts of those who criticize the use of nonbehavioral objectives. Well-chosen illustrations in Chapter 3 relate specific evaluation items to behaviorally defined objectives. The significant influence of evaluation upon student learning is stressed. The discussion of content, predictive, and concurrent validity is related to the instructional functions of

classroom teachers. Construct validity is handled so skillfully as to minimize confusion for the uninitiated.

A conventional presentation of strengths, limitations, and suggestions for writing each kind of item is given in Part II. Well-chosen examples of good and poor items illustrate common pitfalls of item writing. Chapter bibliographies provide excellent sources of sample test-items for various elementary and secondary subject fields. Chapter 9 skillfully develops the use of the interpretative exercise. Text and illustrations combine to elucidate the measurement of such abilities as recognizing the relevance of information; applying principles; recognizing warranted and unwarranted generalizations, assumptions, and inferences; interpreting experimental findings; and utilizing graphs, cartoons, and maps. This emphasis upon objective means of assessing highly transferable learning is much needed.

Part III deals with the use of standardized tests. Readers are spared the lengthy descriptions of instruments that often clutter introductory measurement texts. Students are provided with means of finding, evaluating, selecting, and using the instruments they will need. Transferable skills, rather than specific knowledge, are emphasized.

Two faults appear to mar Gronlund's treatment of Chapter 14, Interpreting Test Scores and Norms. In this well-written but standard explanation of various ways of reporting scores and using reference groups, Gronlund fails to make a clear distinction between the meaning of the word "norm" as it pertains to reference groups and the same word as it pertains to ways of reporting scores—a not uncommon failing. It is the reviewer's unhappy experience that neophites often confuse the two meanings unless the difference is explicitly clarified. The other exception taken with this chapter refers to the statement that "A distinctive feature of percentile norms is that we can interpret a pupil's performance in terms of any group in which he is a member, or desires to become a member" (p. 284). A splendid discussion of the merits of using multiple reference groups follows. One might challenge the assertion that the merits and use of multiple norm groups are either necessarily or logically limited to percentile norms.

In Part IV, Evaluating Procedures, Products, and Typical Behavior, clinical tools are omitted from the discussion, in favor of instruments which typical teachers might use. For instance, the MMPI is not mentioned, and projective techniques are summarized in three paragraphs.

Part V consists of only one chapter. The first part discusses the role of evaluation in improving learning; the second part presents practical and useful suggestions for improving marking and reporting.

In summary, Gronlund has given us an extremely well-written textbook in educational evaluation, the principal theme of which is highly appropriate for introductory courses in educational evaluation and tests and measurements. It achieves its purpose of introducing teachers to the principles and procedures of evaluation which are essential to good teaching.

GERALD S. HANNA
University of Alaska

Psychological Test Modifications by Milton Kornrich (Editor). Springfield, Illinois: Charles C. Thomas, 1965. Pp. xii + 265. $8.75.

Users of projective techniques are frequently faced with the problem of deriving clinical hypotheses from sparse and guarded test protocols. One alternative to this problem in clinical psychology is to devise more objective approaches to personality assessment. This volume, however, takes the point of view that although there is untapped mineral wealth in the testing situation, the usual mining tools require some modifications. Dr. Kornrich has brought together eighteen articles, three previously unpublished, which describe modifications in the test administration of the standard projective techniques to augment the results of the original procedure.

As might be anticipated, most of the articles (seven in all) deal with modifications of the Rorschach. Three articles are devoted to the Thematic Apperception Test (TAT) as well as to measures of word association, two to the Bender-Gestalt, and one apiece to the Children's Apperception Test and drawings. Several of the newer procedures call for eliciting associations to responses obtained in the standard testing session. Perhaps the most unusual method is that of Richard M. Jones' "negated response" technique for use with the TAT. Other procedures involve joint husband and wife testing with the Rorschach and self-interpretation of the TAT.

Several techniques appear to be particularly interesting and perhaps promising. Jones' negated responses and Howard M. Halpern's Rorschach interview technique stand out as procedures that would seem to merit further investigation by clinical psychologists. The usual methodological difficulties in validating projective techniques, however, once again confront the reader.

It would be a simple task to attack the book for being weak in an area not intended to be strong by the editor in the first place. This can be done by pointing out the glaring lack of consideration for validities and reliabilities of the techniques presented. With the exception of one paper by Jones, supporting studies are missing. However, one has to consider the spirit in which the collection is offered. The "dearth of current and past research" is acknowledged in the editor's preface. The purpose is to bring to light some pro-

cedures which may prove helpful in the clinical situation and which merit further study.

Although Dr. Kornrich's collection may offer little to the psychologist concerned with test theory, clinical psychologists may find the book to be lively and provocative reading, particularly if viewed as intriguing suggestions rather than as a clinical manual.

PHILIP HIMELSTEIN
Texas Western College

Widening Horizons in Creativity (The Proceedings of the Fifth Utah Creativity Research Conference) by Calvin Taylor (Editor). New York: John Wiley and Sons, 1964. Pp. xix + 466.

Furnishing a report of the Fifth (1962) Utah Creativity Research Conference held at the secluded and scenic locale of the Mount Majestic Lodge and Manor in the Wasatch Mountains, *Widening Horizons in Creativity* consists of five parts and a total of twenty-nine chapters by a group of distinguished research specialists and writers in the area of creativity. The published papers represent a broad sampling of research contributions from many points of view. Surprisingly enough, even without an overview chapter, there is a marked degree of unity and coherence in the volume that probably reflects the painstaking care exhibited by the editor in his organization of the proceedings.

Titled "Historical Reports," Part I contains two reprinted papers by Toynbee and by Thurstone who were not participants at the Conference. Although written nearly nine years before the Conference was scheduled, Thurstone's paper constituted an amazingly accurate forecast of the problem areas, research methodologies, and substantive contributions to be encountered in later investigations on creative scientific talent. Thurstone's hypotheses were highly insightful forerunners of what is being subjected to systematic inquiry today.

Identified as "Creative Process Studies," Part II consists of seven papers that range from Ghiselin, Rompel, and Taylor's check list of activities associated with the creative process to the clinically oriented paper by Barron on the relationship of ego diffusion to creative perception, to say nothing of Leary's contribution regarding the influence of test score feedback on creative performance and of drugs on creative experience. In addition, there are papers by Westcott on empirical studies about intuition, by the Mednicks on an associative interpretation of the creative process, and by Hyman on the role of information and induced attitudes in creativity. The concluding paper in Part II is a free-wheeling, spontaneous discussion by the Conference participants on process versus product in creativity.

Of considerable interest and value to the professional educator

in the college or university as well as to the public school teacher with adequate psychological background, Part III, "Education and Development of Creativity" is made up of six papers. Torrance reviewed the significant contributions of the Minnesota studies to creative thinking, and Parnes summarized several evaluation studies on research carried out in school settings to develop creative behavior. In his clinically oriented investigation of samples of psychologists whom he designated as creative, noncreative productive, noncreative nonproductive, and combined control groups, Drevdahl noted substantial differences in prior home and educational environments as well as in the current patterns of work habits of these groups. Although not directly concerned with creativity, Harmon's paper on career determiners of high-level personnel yielded validity information on five predictors against a rating of scientific accomplishment. How to develop creative research performance in public school children was the subject of Jablinski's largely anecdotal and impressionistic report. In a related paper Brust discussed the development of readiness in students in the primary and elementary grades for subsequent creative explorations.

Perhaps of greatest relevance to the orientation of the majority of readers of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT was Part IV, "Criterion and Prediction Studies." Taylor and Ellison discussed the prediction of creative performances from use of multiple measures; Guilford expounded on the progress of his systematic research program involving the discovery of intellectual factors; Holland reviewed his work on the assessment and prediction of creative performance of high-aptitude youth; Mullins summarized the findings of four predictive studies involving use of various types of creativity measures; Sprecher described the extent of individual differences in numerous criterion measures for engineers who were being judged for their creative performance; and Astin attempted to conceptualize types of variables for creativity research in terms of high versus low social relevance and naturalistic versus manufactured properties.

For the lay reader as well as for the lay educator, the eight papers in Part V, "Creativity in Special Fields and Settings," probably hold the greatest stimulus value. MacKinnon reviewed his familiar work on creativity in architects; Beittel insightfully considered creativity in the visual arts at the college and university level by showing an amazing grasp of the interrelationship of psychological knowledge and artistic performance; Elliot briefly described the measurement of creative abilities in public relations and advertising work; Fielder outlined a number of studies concerned with the effect of group climate on creative group performance; Gamble summarized the kinds of interest shown and efforts ex-

pended by the National Aeronautics and Space Administration (NASA) in selecting scientists and engineers; McPherson pointed out prospects for future creativity research in industry; Lois-ellin Datta reported on the observations of a committee from industry on the Creativity Conference; and Levine expressed what the efforts and interests were of the National Science Foundation in the development and utilization of creative talent.

Perhaps of as much interest as the main contributions of each of the chapters are the extensive discussions and interchanges to be found within several of the chapters between the speaker (author or co-author of a chapter) and participants at the Conference. These give-and-take sections are not only helpful in clarifying many of the points made by the author or authors of a given chapter but also fertile in suggesting numerous hypotheses about the nature of and the approaches to the study of creativity. Another source of valuable information to the reader is the bibliography of approximately 300 citations.

For the professionally trained psychologist as well as for many a lay person this carefully prepared and edited volume affords a broad overview of the disparate but insightful ideas concerning the processes and products of creative endeavor. As a survey of recent developments in the study of creativity this book probably furnishes one of the most comprehensive and authoritative treatments currently available. After reading this significant book, one can hope that within another four or five years the proceedings of future Utah conferences will be as adequately described as have been those of the Fifth Utah Creativity Research Conference. Hopefully, if a new volume is forthcoming in 1970, its audience will have a basis of assessing and evaluating changes in research efforts on creativity since 1962 and for determining the probable directions that future research efforts may be expected to follow.

William B. Michael
University of California
Santa Barbara

Modes of Thinking in Young Children: A Study of the Creativity-Intelligence Distinction by Michael A. Wallach and Nathan Kogan. New York: Holt, Rinehart and Winston, Inc., 1965. Pp. viii + 357.

This volume contains an investigation categorized in a general sense as systematic observation of the psychological ability called creativity. The authors, Wallach, a member of the Psychology Department at Duke University and Editor of the *Journal of Personality* since 1963 and, Kogan, a Senior Research Psychologist with the Educational Testing Service, have mainly questioned: Can solid evidence be found that will support the validity of a dis-

tinction between the traditional concept of general intelligence and "creativity" as modes of cognitive activity?

Throughout examination of this important research it is essential to keep the operational definition of creativity used in this study well in mind. Mainly the dimension is an associational one, associative flow, the production of associative content that is abundant and that is unique. A further requirement was that the children in the study should generate unique and plentiful associates in a generally task-appropriate context which is permissive and playful. To call a response unique for a given item is to say that the response in question was provided in answer to the item by only one child in the entire sample of 151 fifth-grade pupils. For example, when requested to tell all the different ways to use a newspaper, "rip it up if angry" was a unique response, while "make paper hats" was not (p. 32). For an inverted "v" shaped line in the *Line Meanings* instrument, "private's stripe" was a unique response, but "hat" was not (p. 23).

Concerning the question of whether a response that is unique in the sample as a whole can be considered a creative product for the child, the authors propose that there is a substantial degree of correspondence between "actuarial" and "personal" uniqueness when the sample derives from a reasonably homogeneous sociocultural matrix, such as was the case for the present sample. With regard to the hypothesized concomitance of uniqueness and fluency, in the case of four of the five "creativity" procedures, scored for uniqueness of responses and scored for number of responses, the results were significantly related. Further, the authors expected stereotyped associates to come earlier and unique associates to come later in a sequence of responses. In that case, an appropriate assessment context required freedom from pressure of short time limits, even freedom from any temporal pressure at all, and no evaluative pressure. Consequently children could focus upon the task rather than upon the self. Towards the end of the last chapter, the authors admit that the enumeration of alternative possibilities is but an intermediate step—a precondition for the creative act, and that they do not imply that such a precondition for creativity in a child will necessarily lead to actual creative performance, such as creative problem solving.

The procedures of the study may be summarized briefly as follows. In order to examine the creativity-intelligence distinction and the elucidation of possible psychological correlates of individual differences, seventeen different measures or procedures were administered to the sample of six classes, 70 boys and 81 girls, from two New England schools. These measures (three verbal, two visual) concerned associative flow, cognitive style (band width), emotive connotations (physiognomic properties of stimuli), relatively

enduring motivational sets of defensiveness and anxiety, and intelligence (ten indicators). The measures were reported to be highly reliable—in terms of split-half and item-sum correlations. In future study it would be desirable to know test-retest reliabilities for the "creativity" measures. Besides the seventeen measures, observations were made for a two week period to gather behavioral data on the children and a clinical description was made of thirty-two of the children.

Upon finding that the "creativity" measures were highly intercorrelated (on the order of .4) (an achievement that the authors failed to find in reviewing related research) and upon finding that the correlations between the "creativity" and intelligence measures proved to be extremely low (on the order of .1), the authors then composed four groups of children within each sex: those high in both "creativity" and intelligence, those high in one and low in the other, and those low in both. The number of subjects in each of these cells was approximately 17 or 18. A single "creativity" and "intelligence" score was obtained for each child by summing the standard scores of the measures in the respective domains—possibly a questionable practice in spite of .4 intercorrelations. To yield the groups with the four possible combinations of "creativity" and intelligence, the respective scores were dichotomized at their medians. Next, analysis of variance procedures were used with the other psychological measures to ascertain distinguishing differences.

The following is a sampling of a few of the many findings. Besides the finding of the isolation of an associational dimension of creativity independent of conventionally defined general intelligence, which the authors felt was largely fostered by the permissive and playful test situation, the child's overt behavior in the classroom appeared to vary as a function of his "creativity" and intelligence. For example, although an advantage was found for those who were high in both "creativity" and intelligence, this group was high also in regard to disruptive, attention-seeking behavior (p. 194). In many respects the group high in creativity but low in intelligence was at the greatest disadvantage.

Related to conceptualizing activities, the findings, which may be influenced by ipsative properties of the measurement, focused on the importance of thematizing abilities for "creativity." "Creative" boys seemed able to switch rather flexibly between thematizing and inferential-conceptual bases for grouping (distinctions derived from work by Kagan, Moss, and Sigel). The authors' review of the literature appeared to omit some important studies of the abstract, functional, concrete dimensions, such as, the research by Russell and Saadeh and by Reichard, Schneider, and Rapaport, to mention a few, stemming from developmental theory of Piaget. After their

(possibly inadequate) survey of the literature, the authors rejected the importance of the abstract dimension of cognitive style as related to intelligence and as possibly related to creativity.

In regard to the question of how children set limits or boundaries, the major finding was a relationship between "creativity" and the willingness to tolerate deviant instances as possibly warranting membership in the category, i.e., wider tolerance limits.

To turn to some evidence on how the children describe themselves with respect to anxiety, anxiety level was highest for the low "creativity-low intelligence group. If anxiety was either too low or too high, then "creativity" was reduced. On the other hand, the kind of disturbance represented by defensiveness was not conducive to "creativity" in boys.

Lastly, the ability to produce unique and plentiful ideational associates showed commonality with capacity to apprehend expressive characteristics potentially carried by visual stimuli (physiognomic properties). The case studies tended to support the quantitative results. Explanations for poor performance in one thinking mode, when performance in the other is superior, centered upon motivational interferences.

Perhaps the major contribution of this volume stems from the attempt to design research so that depth of investigation is successfully combined with reasonable breadth, so that both academic psychology and educational research are brought into closer relation, and so that the psychological study of cognitive processes is brought into greater contact with education. At the same time, extreme methodological care and rigor are exhibited in this study. The authors should be commended for attempting to relate their findings to possible school practice (last chapter). Their emphasis, however, is possibly more appropriate for basic or high level dimensions of creative thinking. Perhaps the most significant contribution of this work is that it paves the way for more complex, more relevant aspects of creative thinking to be examined. The authors might have grasped this opportunity themselves with one of their measures originally designed to evaluate defensiveness—a series of unfinished problem-stories, which could have been used to stimulate creative problem-solving behavior in the children. This avenue should be pursued further. Instead, in the reviewer's opinion, most of the "creativity" procedures used in this study smacked of the laboratory game with trivial tasks having little relevance to the life of children. It is hoped, in addition, that future studies will emphasize and build upon knowledge of larger theoretical schemes of children's cognitive development.

The implications from this study for educational and psychological measurement may stem largely from the unique testing atmosphere. It is wondered whether the distinction would be so sharply

maintained if the intelligence measures could have been given in the same playful, permissive, untimed, supportive, atmosphere as the creativity tests were given. It is wondered what effect, if any, results from creativity measures being given in abundance and before intelligence tests.

This study should not be underestimated and should be assimilated and accommodated into present understandings from earlier research on children's creative thinking. In earlier studies claims, too frequently, have exceeded achievements. The findings of this study indicate that "creativity" denotes cognitive functioning that is of importance in the search for new knowledge. The authors have supported the point that to ignore the challenge of measuring particular children with regard to creativity interacting with intelligence is to ignore much of the cognitive potential of young children.

SARA W. LUNDSTEEN
University of California, Santa Barbara

Business Data Processing by Elias M. Awad. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965. Pp. x + 310. $7.55.

Awad's book, the preface indicates, "is designed for a basic course in data processing, for first- or second-year students in schools or colleges of Business Administration. . . . The objective of the book is to provide the student of business administration with a basic and thorough understanding of data-processing principles. . . . Included are principles related to both punched-card and electronic data processing equipment."

In general this text fits the group toward which it is aimed and accomplishes its purpose. This reviewer is particularly impressed by the thorough (perhaps too thorough) treatment of punched-card equipment. Awad has avoided the weakness of many introductory data processing texts: a perfunctory reference to punched-card machines and an overemphasis on the more glamorous computers. He strikes a more suitable balance for a basic, general text—devoting 8 out of 22 chapters to the older "traditional" equipment, and 9 to electronic computers. In rather typical but inescapably logical fashion, Chapters 1 and 2 establish the need for efficient data processing techniques and trace the history of development of computing machinery. In Chapter 1 Awad makes a better than average distinction between punched-card and computer methods of data processing, and presents a better than average explanation of the distinction between digital and analog computers. The historical development of Chapter 2 is interesting and complete.

In Chapters 3 through 10 each of the basic punched-card machines is described and its function explained. Inevitably the reader is thrust into some of the complexities and technicalities of the

workings of the equipment. This section is well illustrated, however, (as is the entire book), and it should provide a good understanding of the applications of punched-card equipment to business data-processing problems.

Chapters 11 through 20 present the basic capabilities which a machine must have to be a computer as well as discuss in somewhat excessive and at times obscure detail some of the mechanical and electronic features of computers. The explanation of the process of reading and writing in magnetic core memory units (pp. 167–168), admittedly technical and difficult to explain, would be unintelligible to the majority of readers. It would be better to omit this "explanation" if it cannot be presented more clearly.

The section on coded decimal schemes (Chapter 17) is of doubtful value, although the topic is almost universally discussed somewhere in other introductory data processing texts. The beginning specialist or general interest reader has no need for this knowledge. In the experience of this reviewer he is usually confused by it.

Many first level data processing texts devote considerable space to the principles of computer programming, sometimes based on a hypothetical machine of the author's, or sometimes using the instruction format (and the inevitable accompanying peculiarities) of an actual computer. Data processing teachers and textbook authors alike are uncertain as to the desirability of including this material in beginning texts. Awad handles the dilemma by briefly presenting (in Chapter 19) a simple example of how programming is done on the IBM 1401, a widely used business computer. Individual instructors could elaborate on this with a more intensive treatment of programming, as could the casual reader, if so inclined, by augmenting this chapter with 1401 programming manuals available from IBM or by reference to a number of texts on programming the 1401.

Chapters 21 and 22 discuss the problems of introducing a computer system into an organization and of managing a computer installation.

This reviewer believes Awad's book is one of the best introductory texts to appear since data processing began to attract the interest of the general public. Although the topic hardly lends itself to easy understanding through an evening's perusal of a book by the fireside, Awad's text is certainly more nearly clear and more suitable for this purpose than many others.

Richard H. Smith
California State Polytechnic College
Pomona, California

The Outcomes of Counseling and Psychotherapy: Theory and Research by Theodore Volsky, Jr., Thomas M. Magoon, Warren

T. Norman, and Donald P. Hoyt. Minneapolis: University of Minnesota Press, 1965. Pp. xi + 209. $5.50.

Since the establishment of the University of Minnesota Testing Bureau in 1932, various groups of educators have devoted their time to the evaluation of counseling, psychotherapy, and casework which occur in different settings with different clients under the sponsorship of a variety of agencies. The present research study was conducted in order to examine the psychological interviews in terms of their outcomes or effects on the clients' behavior instead of in terms of the counseling process. The researchers in this study have attempted to define their problem as being able 1) to identify operationally the goals and/or outcomes of counseling and to determine to what extent such goals and/or outcomes will be realized; (2) to develop criterion instruments which will accurately measure the clients' extra-clinical behavior; and (3) to define as clearly as possible the nature of the counseling process employed in order to achieve the defined objectives.

The general format of this book consists of ten chapters which follow a research outline. A brief summary of the main research topics follows:

In Chapter 1 a general statement of the problems is offered to give direction to the conducting of research in counseling and psychotherapy.

In Chapters 2 through 5 the general methodological topics are reviewed in terms of how they relate to outcome research in these fields. This presentation is by no means complete. However, the authors have selected topics which have often been overlooked in previous counseling studies or which have received inadequate coverage in the current literature. Discussed within these four chapters are the systematic theory formation, design, techniques in analysis, and some associated technical and procedural problems as they relate to outcome research.

Following this discussion on the methodology and procedures, Chapter 6 describes the development of the theoretical framework with specification and definition of key variables which were examined by the authors. This conceptual framework evolved from the counseling or therapeutic processes—objectives to which the counselor subscribes as legitimate, desirable, and expected outcomes of working with clients. The following five dimensions of client behavior were specified and defined: manifested anxiety, defensiveness, personal-problem-solving ability, motivation to change, and perception of counseling.

In Chapter 7 the authors give the details of the development and validation of the criterion measures and classification instruments used.

Chapter 8 outlines the design and analysis as well as the results

of the evaluation of this research. Of the null hypotheses tested by analysis of covariance, only one could be rejected at the .05 level of significance. This rejected null hypothesis indicated that no significant difference existed between the experimental and control groups for the problem-solving variate.

Chapter 9 provides additional descriptive information from two other studies. Jewell's study was conducted because he felt that the direction of the desired movement in manifest anxiety, defensiveness, and solving of personal problems could not be pre-specified. He took the position that these dimensions should be uniquely specified after diagnosis of the individual case. The Vosbeck study investigated the unexamined interrelationships among available sets of data which might provide direction for subsequent research. This study also examined the relationship of client movement on the experimental variables to academic outcomes.

Chapter 10 presents a summary of the conclusions and implications of the research as they may relate to future studies. This chapter is divided into four major issues which the authors hope will stimulate future research in the areas of counseling and psychotherapy. The four areas are: clients and the process, criterion problems, control methods, and future developments.

The authors of this book are to be commended for their insightfulness into the areas of counseling, psychotherapy, and casework. Of particular significance is the discussion of issues and questions which have great importance to researchers conducting studies in these areas. The reviewer feels that this book should be useful especially for those concerned with developing research techniques for future outcome studies and for those concerned with the theoretical work of the neo-behavioral approach to counseling and psychotherapy.

G. Robert Ward
University of California, Santa Barbara

A Social Psychology of Group Processes for Decision-Making by Barry E. Collins and Harold Guetzkow. New York: John Wiley and Sons, 1964. Pp. x + 254.

The purpose of this book is to propose a theory of decision-making in small face-to-face groups. The proposal emerges inductively from the synthesis of selected research findings. The authors are quick to recognize that such a theory, made up of bits and pieces from here and there, is particularly vulnerable to attack, but the authors welcome such attacks, for they recognize that the fruitfulness of their formulations will be measured by the extent to which they are modified by other investigators.

How did Collins and Guetzkow go about synthesizing their theory? First, they abstracted on separate cards those studies which

(a) involved rigorously collected and analyzed empirical and quantitative data, and which (b) were reasonably related to decision-making in small groups. Second, the authors sorted the cards into "topic piles," each with relatively homogeneous themes and results. And third, they formulated propositions based on the grouped material.

The studies which constituted the bits and pieces are drawn almost entirely from small group research. This body of material is broad in the sense that it covers more than decision-making processes in small groups—this by necessity, for little material exists which is specific to decision-making in small groups. But the bibliographic material is narrow in the sense that it is restricted to small groups; there is no review of the literature in the areas of motivation, emotion, learning, or perception, concepts which the authors admit as being important to an understanding of group decision-making. The authors base their "inductive summary and theory" on some 300 small group studies, most of which have been published since 1950. It is natural that the authors rely heavily on the results of the Conference Research project in which Guetzkow, Donald Marquis, Roger Heyns, and others were involved during the years around 1950 at the University of Michigan.

The theory which emerges is expressed in terms of 62 propositions. In addition to its lack of parsimony, the theory sometimes lacks definitiveness. For instance, Proposition 2.4-B states that "The presence of other individuals may increase the defensiveness of the individual; although the effect may be temporary." Of the 62 propositions, 10 are of this type. Another six strike the reviewer as being trivial, and could readily have been written without benefit of a review of the literature. For instance, Proposition 10.3 states that "Interaction with persons we like and persons who like us will produce satisfaction." Still another eight propositions involve tautologies. For instance, Proposition 8.1 states that "High power persons possess more influence," a proposition which must follow from the authors' characterization of "a high power person as one who can influence the behavior of others." Similarly circular, Proposition 3.1 states that "Obstacles originating in the task environment directly inhibit the productivity of individual members," a proposition which is hard to distinguish from the definition on page 70 which states that "A task-environmental obstacle is a particular aspect of the total task environment which blocks, inhibits, or limits group productivity." The proposed concept is inferred from the behavior it is intended to explain.

The authors utilize two units of analysis (the group and the individual in the group) and two classes of independent variables (task-environmental and interpersonal). When the individual is the unit of analysis, it would seem that variables pertaining to the

internal environment (biological variables) would also be important to an understanding of the individual's decision-making behavior. The absence of this important class of variables is a reflection of small group research in general.

Perhaps the most controversial aspect of the Collins and Guetzkow book is their assertion that the concept of leadership is not essential to an understanding of group decision-making, and that the explanatory power of this concept is inferior to that of the more fundamental concepts (common fate, interpersonal attraction, pressure toward uniformity, task-environmental and interpersonal obstacles, task ambiguity, power, satisfaction, and others) of the Collins-Guetzkow theory. However, the authors are more sympathetic to the notion of two leaders within the group, a task-environmental leader (idea person) and an interpersonal leader (best-liked person), a notion consistent with Bales' orientation.

With the conference becoming an increasingly pervasive vehicle for decision-making, there exists an important applied reason for understanding group decision-making processes. Such an understanding will be advanced most rapidly through the conceptualization of the decision-making phenomenon. Collins and Guetzkow have presented such a conceptualization based on the integration of a vast amount of data from diverse sources. Theirs is a book for practitioner and scientist alike.

EDWARD LEVONIAN
University of California, Los Angeles

Integrating the Individual and the Organization by Chris Argyris. New York: John Wiley and Sons, Inc., 1964. Pp. x + 330.

Dr. Argyris has produced a book which should serve as a springboard toward better understanding of organization than has been available in the past. Making use of the same basic facts as have been used in other books on organization theory, Dr. Argyris not only takes into account the high level abstractions usually found in organization discussions, but also focuses on the specific fact of the individual. Recognition that an organization is obviously made up of individuals has been a long time coming.

This reviewer is especially pleased to see how much emphasis was placed on individual energy and how such energy could best be applied toward the effectiveness of the organization. Unfortunately, Dr. Argyris does not have at hand much of the research which has been done in the actual testing of this concept in a large number of organizations. The measurement of the talents of individuals in many organizations has resulted in greatly increasing the knowledge of the interaction which exists in the composition of the total organization as well as in the make-up of all those sub-organizations known as divisions, departments, and similar units.

There are a few questions which arise even in the first few pages of the book. One is the apparent acceptance that physiological and psychological energy are different. There is no reason to assume that these two types of energy are not identical, nor does the creation of "psychological" energy, as such, seem to add to one's understanding of human behavior. A complication has thereby been created which has no essential value in the discussion. Dr. Argyris explains the concept of psychological energy on what appears to be a verbalization of attitudes which reflect the individual's perception of the environmental situation. This creation of concept often results in errors of judgment being made about individual behavior in the work situation. If it is recognized for what it is, a reflection of attitudes, better understanding results. In the world of behavior, these attitudes operating in conflict with the normal output of energy not only result in increased resistance, but also create choking effects on the physiological energy output.

To compensate for the questions created by the energy concepts, Dr. Argyris makes an excellent presentation of the "self" as being an important function of the individual's behavior. Following the understanding of self as being of prime importance, one finds an equal emphasis on the factor of self-esteem and the part which it may play in the success or failure of the individual within the organization. It is further recognized that this individual level of success is reflected in the success or failure of the organization as a whole.

In Chapter 7 the author presents the core of his thinking about organization, in the presentation of the "Mix" model. This model is illustrated as a series of bi-polar concepts, six of which have been spelled out in the book. However, the author recognizes the possibility that other similar concepts may exist and may be essential to the total "Mix" model. The emphasis of the concept of "Mix" is on interrelationships of individuals in the organization and on the awareness of these interrelationships—all directed toward achieving the general objectives of the total organization. It is refreshing that this book does not reject any of the accepted organizational patterns, many of which have been most successful in practice. It does, however, help the reader to recognize that there may be many organizational patterns, all of which may be successful under certain conditions. It is further recognized that these conditions may change and that therefore the pattern of organization may also need to adapt by change.

All in all, this is a most stimulating book, one which needs to be read with an open mind and one which should result in a series of studies to verify the hypotheses suggested by Dr. Argyris. There are two indices of the value of this book to which the reviewer can point. One is that his copy is thoroughly marked emphasizing the

important concepts suggested by Dr. Argyris, and the second is that he intends to use it as a text book in his courses on management theory and organization.

Walter V. Clarke
Providence, Rhode Island

Models, Measurement, and Marketing by Peter Langhoff (Editor). Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965. Pp. 216.

A book of nine essays first presented by the individual authors at the Market Research Council, 1961–62, *Models, Measurement, and Marketing* describes in non-mathematical terms recent advances in model-building, simulation, and decision theory for marketing. Although non-mathematical in exposition, the essays express ideas and concepts which utilize mathematics and the scientific method in application.

Peter Langhoff's essay "The Setting: Some Non-metric Observations" sets the framework for the essays which follow. Langhoff discusses the rapidly developing field of management science and its revolutionary effect upon marketing management and upon model building and measurement.

Philosopher C. West Churchman pursues the pitfalls and problems of model devising and applying models. Churchman whose concern is the conflict between realism and idealism concludes with this basic principle: "a point of view, or a model, is realistic to the extent that it can be adequately interpreted, understood, and accepted by other points of view" (pp. 37–38).

Following Churchman are essays by Harold Kuhn on "Mathematics and Marketing," by Martin Kenneth Starr on "Computers: the Marketing Laboratory," and by Sebastian B. Littauer on "Fundamental Scientific Aspects of Marketing and the Development of Marketing Models." Each of these essays in its own way sets the theoretical framework for model-building, mathematics, and scientific method.

Although each essay specifically relates to marketing, the ideas and concepts advanced in these five articles also have relevance for other areas of the social sciences. For example, the use of computers and Monte Carlo techniques could be used for a predictive purpose for human behavior in general let alone consumer behavior. Also, Professor Kuhn's outline of a curriculum for mathematically training marketing executives is relevant for training hospital or educational administrators as well. Computers can be used for better decision-making by any administrator, and an awareness of mathematics would certainly facilitate their use.

The remaining four essays in the book describe models which have been developed for use in marketing. Alfred A. Kuehn in his

essay "Models for the Budgeting of Advertising" describes three areas of application of models: (1) the determination of the advertising budget, (2) the selection of advertising media, and (3) the evaluation of advertising effectiveness. In "Models of Economic Competition," William J. Baumol approaches marketing decision-making from the point-of view of the economist. Baumol discusses the use of differential calculus, mathematical programming, and game theory for solving business problems. Paul E. Green looks at decision and information theory in his essay on "Decision Theory in Market Planning and Research." Finally Guy-Robert Detlefsen discusses the problem of putting "Theory into Practice."

Perhaps a way to evaluate the book is to look at the objective of the writers in putting the book together. On page 4 Langhoff discusses the objective of the book:

> A major objective of this book is to expose the reader to the basic concepts of mathematical models, information feedback, decision theory, stochastic process, and digital computers and to sketch the opportunities they may have to offer.

To determine for whom the book was written, one reads on the front of the jacket: "A clear exposition of the implications of the new analytical techniques for the work of the marketing executive." However, if the book is designed for the uninformed, the authors in their writings often forget this by escaping into the relative security of technical jargon without acquainting the reader with the meaning of these technical words. In addition, only the last 85 pages are really concerned with the full implication of these techniques for marketing. It is unfortunate that the authors devote relatively more of their efforts to the philosophical bases of the techniques and relatively less to their applications. As a result of this emphasis on theory, the book is deficient of many models which are applicable to marketing: queuing models, warehouse location models, inventory models, and allocation models in general.

The objective of the book is noteworthy, for there is need for the researcher to inform the practitioner of the fruits of his research efforts. But in this instance, the authors spend too much time philosophizing, and except in the last four essays never do come to grips with this problem.

KENT B. MONROE
University of Illinois

Educational Psychology (2nd Edition) by Frederick J. McDonald. Belmont, California: Wadsworth Publishing Company, Inc., 1965. Pp. xviii + 710. $7.50.

McDonald belongs to the new generation of writers in educational psychology who are oriented to the learning emphasis and who are developing systematic theoretical approaches with a learn-

ing model as a structural reference point. McDonald's point of departure is a feedback model as a base for instructional strategies. Nowhere is the break with the older descriptive and developmental orientation more apparent than in the section on Piaget which he introduces with these words, "Piaget's theory . . . is not a learning theory in the same sense that the concept is used in this book. It is a development theory . . . (which) must inevitably be related to theories of learning. However, it has not been to date, nor is it at all obvious how such interrelations can be made . . . Therefore, because this book has taken a learning rather than a developmental approach to instructional problems, we have not studied Piaget to this point" (p. 491).

The new emphasis on learning models and on the systematic development of teaching strategies based on them is fundamental to making teaching an applied science. Each of the models provides the teacher with working hypotheses that can be tested and modified as they are tried in classroom situations. The change is from a folklore to a scientific concept approach.

McDonald's introductory chapters give the student a working knowledge of scientific investigation with the basic statistical tools needed to read with comprehension the research literature. Probably it is inevitable that the first authors who break with a past orientation cannot break completely, or their books will not sell. Much of the middle of the book is quite traditionally descriptive. It is probable that the next edition, and there probably will be many next editions, will pick up some of the points that are troublesome as a strict attempt is made to use the learning model. Although the reviewer is well aware of the fact that curriculum people have differentiated for many years between concept formation and generalizations, the cognitive difference seems to be unreal. Also, the basic relationship between concept formation, developing associations, and memory is not clearly developed.

Another section that probably will be rewritten in the next edition is that on measurement and evaluation. The introductory chapter of this section could probably have assumed that the statistical information furnished in the beginning of the book had been absorbed and this use had been made of those ideas to make the chapter more sophisticated. It seems to the reviewer that the chapter on teacher-made tests is all true and useful material, but is not likely to make any real difference in the kind of tests the readers make for their students. Finally the reviewer was distressed by the chapter on standardized tests. The information on intelligence tests seems to overemphasize the environmental factors and to minimize the variability of scores due to item selection and related hazards of test construction. The emphasis on percentile ranks and ratio IQ's without a word about deviation IQ's or the more common standard score approaches seems unfortunate.

The section on achievement tests is too thin to affect the teaching strategy of any except the most perceptive readers. Teachers make decisions about classroom practice on the basis of their knowledge of the significance of intelligence and achievement scores much more frequently than they do from scores on other standardized tests. In a book for teachers these sections need to be carefully developed. In spite of flaws, this will be a widely used book and one which will grow stronger with successive revisions. The book is beautifully written. The style is clear and lucid. The summary sections are excellent recapitulations of the contents of the chapter. The questions at the end of the chapter are functional for class discussion. A glossary at the end of the book is useful to students. This is an excellent text book for classes that are ready to move toward a theoretical approach to learning as a basis for teaching practice.

John A. R. Wilson
University of California, Santa Barbara

Educational Psychology (Second Edition) by Karl C. Garrison, Albert J. Kingston, and Arthur S. McDonald. New York: Appleton-Century-Crofts Co., 1964. Pp. xvi + 544.

In the revision of their text Garrison, Kingston, and McDonald present an accomplishment indicative of cooperation, integration, and scholarly research. Based upon the recent findings of educational psychology, the data attest that "more than peripheral attention" has been given to current pedagogical problems. The text seeks to remedy the divorce apparent in psychological research through a culmination of data resulting from sound educational theory and practice.

Divided into five major segments the text focuses much attention on the learner. The format of the nineteen chapters, documented by up-to-date research findings and subdivided according to pertinent topic headings, presents a chapter summary, a set of discussion problems, and related readings for extended knowledge.

Of particular interest to readers of Educational and Psychological Measurement is the chapter on "Intelligence," which treats of the usual types of ability tests. Since the Goodenough Draw-A-Man scale (p. 75), which appeared in 1926, has been revised and since many additional data have been presented, it is unfortunate that no indication of this revision is given. (See Dale B. Harris' *Children's Drawings as Measures of Intellectual Maturity: A Revision and Extension of the Goodenough Draw-A-Man Test*. New York: Harcourt, Brace and World, 1963. Pp. 367.) Guilford's "structure of the intellect" and findings on creativity do receive some mention.

In presenting the subtests of the Wechsler Adult Intelligence Scale (WAIS), the writers omit the digit symbol test. In describing samples of tests the authors cite the 1951 edition of the California

Test of Mental Maturity (pp. 334–335). A revision appeared in 1963. A discrepancy appears in the use of the Easel Age Scale recommended for "kindergarten and primary grades" (p. 354). More specifically it is for pupils in kindergarten and grade one according to the test manual. Rothney is cited as the source of the objectives of evaluation (p. 344). They are also to be found in many other early sources—especially in the monumental works of Ralph Tyler and his associates. These are minor points when the text is evaluated *in toto*.

The chapters are not overly general but specific enough to recommend their serious reading and application. Treatment of pupil evaluation embraces test devices both standardized and teacher-made, methods of studying child behavior, diagnostic and remedial treatment of learning difficulties. Discussion of pupil personality along with teacher behavior challenges the reader.

With the up-to-date bibliography, a list of films, and problems for discussion, the text fulfills a definite role in undergraduate and graduate classes in educational psychology.

SISTER JOSEPHINA, C.S.J.
Boston College

Exemplars of Teaching Method by Harry S. Broudy and John R. Palmer. Chicago: Rand McNally and Company, 1965. Pp. vii + 172.

To compensate for the pedestrian nature of many textbooks, the wise teacher will develop a list of outside library readings or other supplementary works, in the hope that at least one of them will "reach" the student and stimulate him to further inquiry. A noteworthy addition to the list is Harry S. Broudy and John R. Palmer's *Exemplars of Teaching Method*. Many readers will perhaps be familiar with Broudy's original essay as it appeared in the *Handbook of Research on Teaching*, edited by N. L. Gage in 1963. The present volume is an enlarged and rewritten version of the original essay. It makes no claim to completeness: its purpose is not that of a textbook in the history of education. The book is intended rather as "a supplement to systematic textbooks in the history of education and as a source of background materials in administration, curriculum, methods, and other standard courses in education."

The book is arranged so as to give one or more eminent examples ("exemplars") of ten different methods and philosophies of teaching, ranging from the ancient rhetoricians to modern progressivists. (The reader will encounter such enticing chapter headings as, "Teaching Courtly Barbarians: Alcuin," and "Dialectical Gardening: Froebel.")

Once the reader identifies with one or more of the exemplars, he

is forced to come to grips with topics that take him far beyond an isolated review of methods used by these teachers. The reader will find himself wrestling with the matters of ultimate objectives, short-term versus long-term goals, the place of vocational, intellectual and physical activities, organization and administration, and a vast range of educational problems. One of the strengths of this approach is that through its sympathetic exposition of diverse teaching methods it provides the student with a whole spectrum of possibilities—possibilities which would otherwise probably never suggest themselves. There is more than one way to carry out the process called education, and it helps to know the advantages and disadvantages of each.

An exposure to this book should cure anyone of the naive view that learning is solely a cognitive matter. The student who unconsciously imitates his teacher is learning on a non-cognitive level how to become a different kind of person. As the authors somewhat rhetorically ask, does the teacher "also serve as an image of a value system that, *aesthetically* apprehended, conveys to the student certain models that invite imitation?" There is no doubt that he does. It might even be the case that the more important things, such as values and attitudes, are taught in this "aesthetic" way.

Let us now turn our attention to several weaknesses or limitations of Broudy and Palmer's approach. The most serious objection might be the lack of adequate coverage. The book does not pretend to offer more than a sampling of exemplars; but even a sampling, to be at all representative, must not limit itself to the western world. What is wanted is a fair selection of models who represent the major traditions in teaching methods. And it just so happens that there are teaching methodologies in the East that have no parallel in the West. Therefore, to limit oneself to the Greco-Roman, Judiac-Christian traditions is to restrict one's educational horizons. There is nothing in the book to even suggest that some of the most interesting methodologies are to be found in the Orient.

Closely connected with the above criticism is the question of the adequacy of the four criteria by which the authors selected exemplars for inclusion in their book. The four criteria are: First, each must have had a "distinctive style of teaching"; second, each must have been "identified with education and formal instruction"; third, besides having a distinctive method, these men also "illustrate a way of life, a system of values, and a response to a cultural challenge"; fourth, they have been "interested in methods and often have been articulate about them. They belonged to the tribe of professional teachers, not only in the sense that teaching was their major life occupation, but also in the sense that the process of teaching and learning seemed to them worthy of serious and specialized study." Religious leaders who were teachers in the wider

sense of the term, such as Jesus or Buddha, were excluded on the grounds that "their goals transcended learning and instruction" and that "they were not conspicuously interested in the methods by which they achieved their results."

It is difficult to see how the Scholastics were selected for inclusion, since their goals surely transcend learning and instruction. Be this as it may, a more serious problem is the exclusion of religious leaders on the grounds that they were not especially interested in the methods used to achieve their goals. But is this true? It is not easy to see how Buddha could be excluded on these grounds, for he systematically expounded and personally practiced one of the most outstanding educational techniques known to man. One can hardly quarrel with the authors' choice to limit the book to secular education if for no other reason than that of space and manageability. But the criteria stated fail to delineate accurately between religious leaders and educators. Perhaps this failure is a clue to the possibility that there is no sharp division between the two.

William Georgiades
University of Southern California

Culture in American Education: Anthropological Approaches to Minority and Dominant Groups in the Schools by Ruth Landes. New York: John Wiley & Sons, 1965. Pp. vi + 330.

Although the nation's attention has been directed by politicians and educators alike to America's "disadvantaged," teachers in the field are frequently perplexed as to how they should personally deal with problems resulting from the diverse subcultures which exist in America. Professor Landes' book is therefore timely and also important as an example of cooperation between anthropology and education.

The book describes the Anthropology and Education program conducted at the Claremont Graduate School from 1959–61. As director of the project Dr. Landes worked with the Claremont education faculty to help Southern California public school teachers understand and to use culture concepts and methods affecting public education. Three main avenues were used: (a) graduate seminars in anthropology and education, (b) classroom observation and discussion with faculty members of neighboring elementary and secondary schools and some junior and four-year colleges, and (c) the development of an elementary school program involving principals and teachers in a large city of Southern California.

The location of the study is significant, for, as the author indicates, California is a microcosm of the world's races, languages, and social conditions. The main minority groups dealt with were Negro and Mexican-American.

Perhaps the most heartening aspect of the difficult task assumed

by the project staff was that teachers learned to understand better themselves and their students by applying basic, simple social science research procedures. Some of the procedures used were: observations of individuals and groups, "open-ended" interviews of informants, and more specific questionnaires. Data collected were shared with fellow students and professors. An immediate objective of the project was for teachers to understand that their own family cultures, like the minorities they were studying, varied greatly. This was accomplished by having each teacher study his own background in a systematic manner.

Results of the Anthropology and Education program are discussed in Chapter Seven, "Teacher Research in Culture and Education." Sections of teacher reports are quoted at length. Thus are demonstrated new understandings gained by teachers as a result of the project's treatment. Professor Landes, in closing, emphasizes the need for effective organization of entire school systems to achieve desired results with active associations between parents and school personnel of primary importance.

Although the book deals specifically with the Claremont Project, it might well be used as a supplementary book in any Social Foundations of Education course. Chapter Three, "Culture Concepts and Methods for Teachers," is an excellent general introduction to culture and its relationship to education. The author brings to this chapter a rich background from a life devoted to the study of anthropology.

Dale L. Brubaker
University of California, Santa Barbara

Handwriting in Psychological Interpretations by Arthur G. Holt. Springfield, Illinois: Charles C. Thomas, 1965. Pp. xiii + 262. $10.50.

The author, a Doctor of Jurisprudence, has devoted forty years of his life gathering the information he offers in this work, and states his case as follows:

> There is a method to acquire extensive knowledge of self and also of others without personal contact with them, which eliminates the danger of being unduly influenced by likes and dislikes. It is the science of handwriting interpretation as demonstrated in this book.

Dr. Holt presents in various styles of writing "standard letters of domestic and foreign alphabets" and comments upon each letter as to the character traits and personality tendencies of the writer. Preliminaries are also offered for further analysis along with other pertinent information.

Chapter 2, entitled "The Science of Graphology and the Underlying Theory of This Approach," contains a brief history of the

subject. The reader is told: "In America many research materials are available, but very few are written in textbook form." Also, the reader is appraised of the fact that West German and French Universities include graphology in a "clinical or applied psychology curriculum."

What the author does do effectively is to state his belief clearly and to present descriptive information. What is needed now, however, is carefully researched experimental data. Perhaps another volume containing material of this nature will follow.

ARTHUR LERNER
Los Angeles City College

The Underachiever: Challenges and Guidelines by C. Burleigh Wellington and Jean Wellington. Chicago: Rand McNally & Company, 1965. Pp. x + 122. $1.75.

This paperback contains a wealth of information about the underachiever, including existing research and background material for identification, selection, and assessment problems in the field. The author believes that through an understanding of the similarities among underachievers sound helping programs can be established. Although the book is essentially oriented to the school situation, much of the information may also be helpful to the parents of underachievers.

The Underachiever: Challenges and Guidelines, which makes a serious attempt to explain research findings in a non-technical manner, will prove to be a handy reference on many practical and theoretical levels. Its presence is a vital contribution to a most complex phenomenon of the educational experience.

ARTHUR LERNER
Los Angeles City College

Educational and Psychological Measurement
1966, 26, 545-564.

MULTIPLE FACTOR ANALYSIS IN TERMS OF WEIGHTED REGRESSION[1]

JOHN M. BUTLER and L. HARMON HOOK
The University of Chicago

Some basic problems in factor analysis have been clarified in the last two decades in important papers by Guttman (1940, 1953, 1956), Lawley (1940), Rao (1955), Harris (1962, 1963) and Kaiser[2] (1958, 1963). In particular their work clarifies the concept of what Thurstone (1947) called unique variance and the communality concept or closely related concepts. It is, however, the very centering upon the communality problem and closely related problems which constitutes a consequential departure from multiple factor analysis as envisaged by Thurstone. As Thurstone envisaged it, factor analysis should have as a main purpose specification of the domain or population of factors. For any given sample of tests it might be that population common factors are in varying degrees well represented, poorly represented or not represented at all. Common factors poorly and fractionally represented in the tests were called specific factors by Thurstone, and he stated that such factors may become common factors in a test battery when a given battery is augmented by additional tests. It follows from what Thurstone has said about specific factors that the factorial domain cannot be specified until a test battery well represents that domain. Since in practice the domain is known through a set of related measurements, it seems somewhat beside the point to estimate

[1] The views presented here are an outgrowth of research supported by research grants USPHS MH 04609-04, USPHS MH 10131-01, NIMH, U. S. Public Health Service.

[2] The writers have been considerably influenced by the papers of these authors. However, the influence is so pervasive that little attempt will be made to particularize the indebtedness beyond the listing of references.

population communalities or uniquenesses unless one has good reason to believe that a test battery is representative of a factorial domain. In this case there would be no specific variance in the test battery. Unfortunately, zero specific variance cannot insure that a factorial domain is well represented in a test battery because of the possibility that some factors may not be represented at all. For these reasons the discussion here will be confined to the sample of tests, and the communality problem and the problem of the number of factors will be considered in relation to sets of vectors which will be designated as unique vectors and specific vectors or factors.

Harris (1963) and Guttman (1956) have obtained uniqueness values, v_j^2, such that

$$v_j^2 = [r^{jj}] \cdot [(r^{jk})^2]^{-1} \tag{1}$$

where r^{jj} is a row vector composed of the diagonal elements of R^{-1}, the inverse of the correlation matrix, and $(r^{jk})^2$ is the square matrix with elements consisting of the squares of the corresponding elements of R^{-1}. These uniqueness values have the disadvantage that they may be greater than maximum uniqueness values; furthermore, they may assume negative values (Harris, 1963, p. 147). Evidently, then, the v_j^2 do not conform to the fundamental requirements that uniqueness values be positive and that communalities have a maximum value of unity.

The Factor Model

The approach to be taken here will be based upon regarding the uniqueness–communality problem as a problem in weighting a regression equation. This amounts, as will be shown later, to weighting maximum (and always positive) uniqueness values by positive values between zero and unity.

Let S be the $n \times t$ matrix of normalized standard scores (Thurstone, 1947, p. 367) for n tests and t subjects with each test vector being orthogonal to no more than $n - 2$ test vectors. Then let F be an $n \times n$ matrix such that

$$SS' = FF' = R \tag{2}$$

where R is the correlation matrix. Then F is a matrix representation of the test vectors in an arbitrary orthogonal basis such that the row vectors, as representations of the test vectors have unit

length. F, being square and having rank n, has an inverse, F^{-1}, so

$$FF^{-1} = I. \tag{3}$$

Equation (3) shows that the identity matrix represents the perpendicular projections of the column vectors of F^{-1} upon the unit length row vectors represented in F. Since each column vector of F^{-1} is orthogonal to every row vector of F but one, the column vectors of F^{-1} are the unique vectors for the sample of tests. Normalizing the column vectors by post-multiplying F^{-1} by the diagonal matrix of normalizing constants, D, gives

$$FF^{-1}D = D. \tag{4}$$

Equation (4) shows that D is the matrix of intercorrelations of the unique and test vectors.

Given the elements of D as the correlations between the test vectors and the unique vectors it has been shown (Guttman, 1940) that

$$I - D^2 = \bar{R}^2 \tag{5}$$

where $\bar{R}$ is the diagonal matrix of multiple correlations between each test and the remaining $n - 1$ tests. $\bar{R}^2$ is the diagonal matrix of communalities corresponding to the test vectors when the unique vectors have unit length. These are the minimum communalities (Guttman, 1953, 1956). It should be noted also that the elements of D^2 are maximum values because the unique vectors cannot be taken as having length greater than unity. The maximum uniquenesses and minimum communalities are thus functions of cosines of angles.

In multiple factor analysis the test vectors may be considered to be the vector sum of the unique vectors and what will here be called "common" vectors. The equation for the vector sum, expressed in terms of F is

$$F = C + KF'^{-1} \tag{6}$$

where C is the matrix of common vectors and KF'^{-1} is the matrix of unique vectors with the unique vectors being adjusted for length by the diagonal matrix K with elements k_i such that

$$k_i \leq d_i^2$$

where d_i^2 is the i-th element of D^2. The row vectors of F, the matrix of test vectors; C, the matrix of common vector s, and KF'^{-1}, the

matrix of unique vectors, are all expressed in the same arbitrary orthogonal basis.

Now let $K = D^2$ and $C = WF$ where W is a matrix of weights to be applied to F to obtain C. Then (6) becomes, after some rearranging,

$$C = WF = F - D^2F'^{-1}. \tag{7}$$

Post-multiplying (7) by F^{-1} gives

$$W = I - D^2F'^{-1}F^{-1}. \tag{8}$$

It is known (Guttman, 1940; Kaiser, 1963) that W is the matrix of multiple regression coefficients to be used for predicting a given test score from the remaining $n - 1$ tests. Thus the common vectors are the predicted vectors when a least squares criterion is used. Note that the diagonal entries of D^2 are the inverses of the diagonals of $F'^{-1}F^{-1}$ so the diagonals of W are zero. What (8) shows is that the use of D^2 for uniqueness values is equivalent to regarding factor analysis in multiple regression terms (D^2 is the error variance in the regression equation).

C may also be obtained quite simply by subtraction. However, when this is done K is not specified.

F may be transformed to $GH^{1/2}$ where G is an orthogonal matrix and $H^{1/2}$ is a diagonal matrix such that

$$GHG' = R. \tag{9}$$

Then $GH^{1/2}$ is a principal axis factor matrix for the correlations and the matrix of common vectors can be written in principal axis terms as follows:

$$C = GH^{1/2} - D^2GH^{-1/2}. \tag{10}$$

Pre-multiplying C by D^{-1} gives

$$D^{-1}C = D^{-1}GH^{1/2} - DGH^{-1/2}. \tag{11}$$

Since $G_iH^{1/2}$ and $G_iH^{-1/2}$ represent corresponding test and unique vectors, the inner products of the vectors, $D_i^{-1}G_iH^{1/2}$ and $D_iG_iH^{-1/2}$ are unity, and those of $D_i^{-1}G_iH^{1/2}$ and $D_iG_iH^{-1/2}$ are zero.

From the right side of (11) it can be seen that corresponding to $D^{-1}GH^{1/2}$, the weighted principal axis matrix, is another principal axis matrix, $EP^{1/2}$, and corresponding to $DGH^{-1/2}$ is another principal axis matrix $EP^{-1/2}$. $EP^{1/2}$ can be obtained from $D^{-1}GHG'D^{-1}$.

(11) can then be rewritten in terms of $EP^{1/2}$ and $EP^{-1/2}$ as follows:

$$\begin{aligned} D^{-1}C &= EP^{1/2} - EP^{-1/2} \\ &= E(P^{1/2} - P^{-1/2}). \end{aligned} \tag{12}$$

The covariance matrix for the weighted common vectors is

$$E(P^{1/2} - P^{-1/2})^2 E'. \tag{13}$$

It is important to note that factoring the covariance matrix of the weighted common vectors (13) to obtain the principal axis factor matrix will not give $E(P^{1/2} - P^{-1/2})$ except when $P_i^{1/2} > 1$. Thus a proper factoring of the covariance matrix of the weighted common vectors is given by $E(P^{1/2} - P^{-1/2})$ and *not* by $E[(P^{1/2} - P^{-1/2})^2]^{1/2}$ except when all values of $P^{1/2}$ are greater than unity.[3]

The weighted test vectors and common vectors can be returned to their original scale by pre-multiplication as follows: $DEP^{1/2}$ is a factor matrix for the test vectors and $DE(P^{1/2} - P^{-1/2})$ is a factor matrix for the common vectors.

The column vectors of $DEP^{1/2}$ and $DE(P^{1/2} - P^{-1/2})$ are the same except for scale for those values of $P^{1/2}$ above unity but not for those values of $P^{1/2}$ less than unity. This conclusion may be restated as follows: Given the covariance matrices of the weighted common vectors and the weighted test vectors and applying principal axis factoring routines results in factor matrices congruent to I_n whereas obtaining $E(P^{1/2} - P^{-1/2})$ directly from $EP^{1/2}$ gives a factor matrix such that $E(P^{1/2} - P^{-1/2})E'$ is congruent to diag. $(I_p, -I_{n-p})$ where p is the index integer. Thus only the first p factors are common factors in the total common factor space of the sample of tests. The index integer p is the same number as Guttman's strongest lower bound for the number of factors in the population.

The preceding development is based on a regression point of view in which D^2 represents unpredictability. The logic of this point of view requires that the entire true variance be common factor variance while the logic of the multiple factor analytic point of view enunciated by Thurstone requires that the elements of D^2 represent upper bounds to the uniquenesses, being the uniquenesses only when the variance due to specific vectors is zero. Now specific

---

[3] Guttman (1953) and Kaiser (1963) define the partial image vectors in terms of the matrix $DE[(P^{1/2} - P^{-1/2})^2]^{1/2}$, so common vectors and partial image vectors are not in general identical.

vector variance is common factor variance from factors poorly and fractionally represented in the tests of the battery (Thurstone, 1947, p. 75), so poorly and fractionally that a regression approach counts it as error variance since the regression model, which is addressed to prediction of each test from the remaining tests, is concerned only with maximizing predictability. Specific variance is a sample phenomenon and is relative to a particular battery of tests. What is specific variance in a given test battery may become common factor variance when the battery of tests is augmented by other tests. It follows that when a sample of tests does reflect a population of factors adequately, there will be no specific variance. On the other hand, it does not follow that the test battery adequately reflects the domain when there is no specific variance because the battery may not represent some of the factors in the domain at all. Guttman (1953, 1956) has stressed that the side entries of the inverse of the actual correlation matrix must be "small" or zero if the data are to conform to the factor analytic model or if the test battery is to be considered representative of the universe of content (or of the population of factors). The requirement that the inverse of the correlation matrix be diagonal or "nearly" diagonal and the assertion that the test battery then adequately taps the population of factors leads logically to the requirement that the squared multiple correlation coefficient be used at least initially as a communality estimate. Then using additional tests to augment a test battery can only lead to changes in squared multiple correlations occasioned by error variance.

Now all of this is true enough if one knows or has good reason to believe that the battery of tests adequately reflects the population of factors. However, when one is using factor analysis as an exploratory device in the manner of Thurstone, one does not have good reasons for believing that the tests adequately tap the entire population of factors. The investigator may hope that this is so but there is no assurance that it is so. Furthermore, psychological tests do not have independent existence of their own; they are constructed on the basis of psychological hypotheses. They have a quite different status than natural objects such as corn, swine, and skulls or other objects of statistical analysis and interest such as definite products of manufacture like nuts, bolts, and light bulbs. For these objects face validity corresponds to actual validity

whereas for psychological tests a test with high face validity may not at all perform the task for which it was designed. *Since tests are constructed on the basis of psychological hypotheses, the "population" of tests may well depend only on the fertility of the human mind and populations may be continually redefined.* That additional test construction will *not* result in new factors in any domain so far investigated seems fairly unlikely. It seems, therefore, that in the absence of compelling reasons for believing that a test battery is representative of a definite population of tests (or of factors) it is somewhat feckless to think of population estimates of communalities or of uniquenesses. The situations in which the population of factors is known or in which the sample of tests is representative are, to understate the case, rare.

Sample Specific Factors

A sample specific vector must be uncorrelated with the remaining $n - 1$ specific vectors. The i-th specific vector must have as high correlations as possible with the remaining test and unique vectors. The i-th specific vector can have zero correlations with the j-th test and unique vectors only when the inverse of the correlation matrix is diagonal. Letting D_f^2 denote the diagonal matrix of squared correlations of the i-th specific and the i-th unique vectors we have the proportion of the i-th unique vector attributable to specific factor variance. The squared correlations of the diagonal matrix D^2 give the proportions of the test variances which are unique variance, and the matrix product $D_f^2D^2$ gives the proportion of the total test variances attributable to specific factor variance. That is, they give the proportion of common factor variance poorly and fractionally represented in the zero-order correlations of the correlation matrix.

Given D_f^2 the factor matrix for the unique vectors can be weighted by D_f^2 to get

$$F = C + D_f^2D^2F'^{-1}. \tag{13}$$

(13) expresses the regression equation weighted by the proportion of the unpredictability which is occasioned by factors poorly represented in the tests. Since the elements of D_f are correlation coefficients, the elements of $D_f^2D^2$ can be unity only when D^2 is an identity matrix and D^2 will be an identity matrix only when R^{-1} is diagonal.

The elements of the weighted uniqueness matrix have the following desirable properties not possessed by the Guttman-Harris uniqueness values (Harris, 1963) and required by the multiple factor analytic model:

1. Elements of $D_f^2D^2$ are always positive.
2. Elements of $D_f^2D^2$ are always in the range from zero to d_i^2 where d_i^2 is a diagonal element of D^2 and $0 < d_i^2 < 1$.

It seems likely that the elements of $D_f^2D^2$ will each be greater than any "true" uniqueness values.

The matrix of weighted maximum uniqueness values has not been specified in terms which admit of computation. However, the set of elements for D_f^2 can be specified as a consequence of noting that the i-th specific vector must have higher correlations with the i-th test and unique vectors than with the j-th test and unique vectors. Consider the matrix,

$$L(F + DF'^{-1}) \tag{14}$$

where L is a diagonal matrix of normalizing constants for the row vectors of $F + DF'^{-1}$, and the matrices of correlations

$$L(F + DF'^{-1})F' = L(R + D) \tag{15}$$

and

$$L(F + DF'^{-1})F^{-1}D = L(DR^{-1}D + D). \tag{16}$$

Since the elements of L are less than unity, it is clear that the correspondingly placed diagonal correlations of (15) and (16) are equal and that the off-diagonal correlations of (15) and (16) are less than the correspondingly placed correlations in R and $DR^{-1}D$. Also the off-diagonal elements of the right side of (16) are in general less than those on the right side of (15). These facts suggest that the row vectors of (14) or of suitable linear combinations of them can be used to obtain suitable values for D_f^2; i.e., the row vectors of (14) or suitable combinations of them can appropriately be viewed as specific vectors. Although the diagonal correlations of (16) might be used as values of D_f it is generally true that the row vectors of (14) are not orthogonal even though (14) more nearly approximates an orthogonal matrix than does F or DF'^{-1}. However, an orthogonal matrix with row vectors possessing the salient characteristics of the specific vectors may be obtained by iterating on

the process employed in obtaining (14). The iteration process is described below.[4]

Let F_0 be the $n \times n$ matrix with test vectors as rows such that

$$F_oF_o' = R_o,$$

the correlation matrix. Then $D_0F'_0{}^{-1}$, where D_0 is the diagonal matrix of normalizing constants for the row vectors of $F'_0{}^{-1}$, is a factor matrix for the matrix of intercorrelations of the unique vectors and

$$F_oF_o^{-1}D_o = D_o \tag{17}$$

is the matrix of correlations of the test vectors and the unique vectors.

The unique vectors as rows of $D_0F'_0{}^{-1}$ are represented in the same orthogonal basis as the test vectors. The centroid vectors formed by summing corresponding test and unique vectors comprise the row vectors of G_1 where

$$G_1 = F_o + D_oF'_o{}^{-1}. \tag{18}$$

Normalizing the rows of (18) by the diagonal matrix, L_0, gives

$$F_1 = L_oF_o + L_oD_oF'_o{}^{-1}. \tag{19}$$

The matrix of correlations of the row vectors of (19) is

$$R_1 = F_1F'_1. \tag{20}$$

R_1 more nearly approximates an identity matrix than does R_0 since the row vectors of F_1 are the unit centroid vectors formed from corresponding test and unique vectors. Since the rows of F_1 are unit length vectors, F_1 more nearly approximates an orthogonal matrix than does F_0. F_1 is non-singular and has an inverse. Iterating upon the procedure used on F_0 gives

$$F_2 = L_1F_1 + L_1D_1F'_1{}^{-1} \tag{21}$$

where L_1 is the diagonal matrix of normalizing constants for the rows of $F_1 + D_1F_1^{-1}$ and D_1 is the diagonal matrix of normalizing constants for the rows of $F'_1{}^{-1}$.

Since

$$F_2F_2^{-1} = I, \tag{22}$$

[4] Although there may be a theoretically better solution based on using latent vectors rather than centroid vectors, the present approach seems satisfactory and should in any event yield results virtually the same as those obtained from using latent vectors.

the row vectors of F_2 bear the same relationship to the row vectors of $F'_2{}^{-1}$ as those of F_0 bear to $F'_0{}^{-1}$.

Hook (Butler, *et al.*, 1963, App. B) has shown that the sequence of matrices

$$R_0, R_1, R_2, \cdots, R_f \tag{23}$$

formed as above converges to the identity matrix. This convergence in turn implies that the sequence of matrices, F_i with unit length row vectors, all vectors being expressed in the same orthogonal basis,

$$F_0, F_1, F_2, \cdots, F_f \tag{24}$$

converges to an orthogonal matrix (but not to an identity matrix). The matrices

$$F'_1F_1, F'_2F_2, \cdots, F'_fF_f \tag{25}$$

give the inner products of the column vectors of the F_i. F'_fF_f approximates an identity matrix as closely as one wishes.

The matrix of correlations of the unique vectors and the vectors of the final orthogonal basis of unit centroid vectors is

$$\hat{F}_f = D_0F'_0{}^{-1}F'_f. \tag{26}$$

The i-th unit centroid vector represented by the i-th row vector of F_1 is determined entirely by the i-th row vector of F_0, the test vectors in an arbitrary orthogonal basis, and the i-th row vector of $D_0F'_0{}^{-1}$, the matrix of unique vectors expressed in the same orthogonal basis. Note, however, that each unique vector is actually determined by all of the $n - 1$ test vectors except its corresponding test vector.

Thus, the i-th centroid vector bisects the angle separating the i-th test and unique vectors and is, therefore, the vector simultaneously most similar to both the i-th unique and test vectors. It will have relatively high cosines with the i-th test and unique vectors and relatively low cosines with the remaining $n - 1$ test and unique vectors. The unit centroid vectors of F_1 comprise a collection of vectors more nearly approximating orthogonality than the collection of unique vectors, but, as with the unique vectors, the i-th centroid vector will be similar to the i-th test vector and dissimilar to the j-th test vector.

The diagonal matrix, D, is a diagonal matrix of correlations for

weighting unique vectors and the diagonal values of $\hat{F}_j$ are the correlations of the unique vectors with their corresponding specific vectors. If d_i is high, then the corresponding diagonal value of $\hat{F}_j$ will be high. Conversely, if d_i is low, then the corresponding diagonal element of F_j will be low. The matrix of diagonal values of F_j;

$$\text{diag. } \hat{F}_j = D_j \tag{27}$$

of correlations of uniqueness vectors with corresponding specific vectors gives the similarity of each uniqueness vector with its corresponding specific vector. It should be noted that, from their manner of formation, the diagonal values of F_j will be equal to or greater than the off-diagonal values.

The use of D_j^2 as a weight matrix additional to D^2 utilizes the fact that starting with F_0 and F_1 and taking successive sets of unit centroid vectors simultaneously best fitting corresponding test and unique vectors, a final set of orthonormal specific vectors is obtained having cosines as high as possible with corresponding test and unique vectors and cosines as low as possible with the remaining $n - 1$ test and unique vectors.

The equation corresponding to the modification of (10) by D^{-1} is

$$D_j^{-1}D^{-1}C = D_j^{-1}D^{-1}GH^{1/2} - D_jDGH^{-1/2}. \tag{28}$$

The corresponding row vectors of the matrices on the right of (28) have unit inner products and the non-corresponding row vectors have zero inner products. The principal axis matrices corresponding to the matrices of (28) may, as before, be written in the form

$$E(P^{1/2} - P^{-1/2}), \tag{29}$$

$$EP^{1/2}, \tag{30}$$

and

$$EP^{-1/2}, \tag{31}$$

although the latent vectors and latent roots are different from those corresponding to weighting by D^2 only.

Once $D_j^2D^2$ has been obtained for use as a weight matrix, the weighted correlation matrix

$$D_j^{-1}D^{-1}RD^{-1}D_j^{-1} \tag{32}$$

should be factored by a principal axis factoring to obtain $EP^{1/2}$.

No more than p factors should be extracted where p is the index integer corresponding to $E(P^{1/2} - P^{-1/2})E'$. This amounts to taking as the principal axis factor matrix of the common vectors those rows of $E(P^{1/2} - P^{-1/2})$ for which the values of $(P^{1/2} - P^{-1/2})$ are positive. This gives the maximum number of factors in the sample common factor space for a given $D_i^2D^2$.

Number of Factors To Be Taken from the Sample

The practice of taking a much greater number of factors than has been usual among factor analysts has been advocated by Guttman (1953), Harris (1963) and Kaiser (1963). There is reason, however, for viewing this practice, which seems to be becoming conventional, with some reserve. First: A consideration of the number of degrees of freedom available for testing the number of factors shows that the number available is a function of the number of tests and the number of factors extracted. Suppose the number of factors is taken as p, the number of positive values of $(P^{1/2}—P^{-1/2})$. This is the index number of diag. $(I_p, —I_{n-p})$ with index integer p, and $n—p = r$, the signature integer, where n is the number of tests. Then the number of degrees of freedom available for testing the number of factors (Rao, 1955) expressed in terms of the index and signature is

$$d.f. = [r^2 - (n + p)]/2 \tag{33}$$

which shows that the inequality

$$r^2 > n + p \tag{34}$$

must hold if degrees of freedom are to be available for testing the hypothesis that there are at least p common factors in the population. With an n of 12, a p of 8, and an r^2 of 16, $n + p$ is 20. Hence the maximum testable number of factors for a battery of 12 tests is 7, even though Guttman's strongest lower bound for the number of factors in the population may exceed 7. Thus the practice of extracting p factors cannot be applied without consideration of the number of tests in the battery. Of course, the condition that (34) hold will be met as the test battery increases in size. However that may be, it does not alter the fact of the general inapplicability of the practice.

Second: Consider the inner products of the weighted test vectors, represented in the rows of $EP^{1/2}$, and the row vectors of $E(P^{1/2} — P^{-1/2})^{-1}$. These inner products are the elements of

$$EP(P - I)^{-1}E'. \tag{35}$$

The row vectors of $E(P^{1/2} - P^{-1/2})^{-1}$ should have small inner products with the test vectors of $EP^{1/2}$. Table 1 shows that these inner products will be small only when $P_i \geqq 2$. When this is so, the latent roots of (35) decrease from 2 to 1.01 as the latest roots, P_i, increase from 2 to 100. As P_i goes from 2 to 1, however, the latent roots of (33) go from 2 to 101. Thus as P_i decreases below 2, the elements of $EP(P - I)^{-1}E'$ increase rapidly in magnitude.

TABLE 1

Values of Various Latent Roots

P_i	$P_i(P_i - 1)^{-1}$
100.00	1.01
50.00	1.02
20.00	1.05
10.00	1.11
9.00	1.12
6.00	1.20
5.00	1.25
4.00	1.33
3.00	1.50
2.50	1.67
2.00	2.00
1.50	3.00
1.01	101.00
1.00	...

This implies a decreasing similarity of the weighted test vectors to the weighted common vectors as the P_i decrease. It also implies, that when P_i^{-1} are below unity, rotation to simple structure will eventuate in null or uninterpretable factors, the number of such factors being the number of latent roots, P_i^{-1}, with values below unity. This is shown directly by the matrix of inner products of the weighted test vectors and the weighted common vectors,

$$EP^{1/2}(P^{1/2} - P^{-1/2})E' = E(P - I)E'. \tag{36}$$

The greater the (P_i^{-1}), the greater the off-diagonal elements of (36) will be.

It should be noted that pre- and post-multiplying (36) by DD_i and $(DD_i)'$ gives

$$DD_iD_i^{-1}D^{-1}RD^{-1}D_i^{-1}D_iD - D_i^2D^2 = R - D_i^2D^2, \tag{37}$$

the correlation matrix with the communalities in the diagonals.

Since $E(P - I)E'$, like $E(P^{1/2} - P^{-1/2})E'$, is congruent to diag. $(I_p, -I_{n-p})$, and since the diagonal values of $(P-I)^{1/2} > (P^{1/2}-P^{-1/2})$, the inner products of

$$(E_iP^{1/2})(P - I)^{1/2}E'_i > (E_iP^{1/2})(P^{1/2} - P^{-1/2})E'_i.$$

Hence, $E(P - I)^{1/2}$ is a better factor matrix than $E(P^{1/2} - P^{-1/2})$ because of higher inner products with the original variables and, furthermore, using $E(P - I)^{1/2}$ is equivalent to factoring the correlation matrix with communalities in the diagonals for $D_i^{-1}D^{-1}E(P - I)^{1/2}$ is a weighted principal axis factor matrix for the correlation matrix with communalities in the diagonals (Harris, 1962, 1963).

The results of Table 1 suggest that the number of factors to be extracted from the correlation matrix should be restricted to the number of values of P, the diagonal matrix of latent roots of the weighted correlation matrix,

$$D_i^{-1}D^{-1}RD^{-1}D_i^{-1} \tag{38}$$

equal to or above 2. The maximum number of factors will be the number of values, P_i, above unity. Further (34) and (35) indicate the general inapplicability of the procedure of extracting p factors.

Finally, and crucially, observe that R can be written as follows:

$$R = FGF' - I + I \tag{39}$$

and

$$R = F(G - I)F' + I \tag{40}$$

so we have R and $R - I$ expressed in terms of a common set of unit latent vectors and the latent roots of G and $G - I$.

With the restrictions earlier placed on the correlation matrix, rank n and non-zero off-diagonal correlations in each row, there will be at least one latent root below unity in R. It follows that $R - I$ is in the same congruence class as J where

$$J = \text{Diag.}\,(I_r, -I_{n-r}) \tag{41}$$

with index integer r. Let X be any diagonal matrix with positive constants. Then $X(R - I)X$ is also a member of the canonical class J. Thus the number of factors determined solely by the side entries of the correlation matrix is the index integer r of J, which is

also the number of latent roots of the correlation matrix above unity.

Steps in Factoring

For practical purposes it is recommended that:

1. The number of factors be taken as the number of latent roots of the correlation matrix above unity.
2. Use $D_f DE\ (P - I)^{1/2}$ as the final prerotation factor matrix.
3. Rotate the factors of (2) to an orthogonal approximation to simple structure (Kaiser, 1963).
4. If s essentially null factors appear upon rotation take $r - s$ factors of $D_f DE(P - I)^{1/2}$ and re-rotate. Continue the process until one essentially null factor appears after rotation.
5. If an oblique simple structure solution is desired, rotate only after first obtaining the orthogonal approximation to simple structure (Kaiser, 1963).

These recommendations bearing on rotation were written on the assumption of use of analytical means of rotation. If graphical rotation is resorted to, then direct rotation to oblique simple structure can be accomplished. Simple structure should be the final arbiter of the number of factors in a sample, not statistical tests.

Final Factor Matrices

The correlations between the first p factors of

$$EP^{1/2}$$

and

$$E(P - I)^{1/2}$$

are unity. Therefore, although $E(P - I)^{1/2}$ is equivalent to a factoring of the correlation matrix with communalities in the diagonals, one could elect to use $EP^{1/2}$ as the matrix from which to rotate the simple structure for both $EP^{1/2}$ and $E(P - I)^{1/2}$ represent different weightings of the correlation matrix by functions of the communalities.

There is reason to believe that for a given test battery $EP^{1/2}$ the weighted test vectors might have a somewhat better simple structure than the weighted test vectors of $E(P - I)^{1/2}$. The latter matrix is obtained by subtracting unity from each latent root of P_i. The result is that the last latent roots of $P - I$ become dispropor-

tionately small in relation to the first principal axis when compared with the latent roots of P. For example,

$$P_1 \Big/ \sum_1^{r-1} P_j < P_1 - I \Big/ \sum_1^{r-1} P_j$$

where $1 \neq j$ and r is the number of factors. This means that in using an analytic solution such as normal varimax, the simple structure factor with the highest cosine with the first principal axis is likely to have higher factor loadings for $E(P - I)^{1/2}$ than for $EP^{1/2}$. It should be remembered, however, that the number of factors to be extracted must be the number of latent roots of the correlation matrix above unity if one wishes to restrict his factors to the common factor space of the side correlations of the correlation matrix.

Table 2 gives final factor matrices for a set of body measurements (Thurstone, 1946). The correlation matrix had three latent roots above unity. However, the fourth latent root was close to unity so four factors were extracted. The factor matrices seem to be identifying the same simple structure. For practical purposes the simple structures in this case are virtually identical. The factor matrices are based on the use of D_1, obtained from F_1 (19), as an approximation to D_j.[5]

Comments, Conclusions and Summary

The multiple factor analytic factor model represents intuitive ideal conceptions of common, unique, specific and error vectors. Vectors exist in the sample corresponding to these concepts. The concept of specific vectors is of peculiar importance for factor analysis conceived as a tool for exploratory purposes, as a device for aiding an investigator to define and re-define a domain. Specific variance represents the acknowledgment and the assertion that part of what is unpredictable is real and the refusal to take refuge in the concept of what is in the nature of the case an indefinable population of factors. The specific vectors constitute a definition of the unrepresentativeness of tests with respect to populations of factors. When a battery of tests is representative of the population of factors, specific variance does not exist. However, a battery of

[5] Computer program UCSL 505 for obtaining $D_1^{-1}D^{-1}RD_1^{-1}D^{-1}$ is in the files of the Social Science Program Library of the Computation Center of The University of Chicago.

TABLE 2

Alternative Orthogonal Simple Structure Matrices Based on $EP^{1/2}$ and $E(P-I)^{1/2}$

	Simple Structure From $D_1DEP^{1/2}$					Simple Structure From $D_1DE(P-I)^{1/2}$						Thurstone's Factor Matrix				
		I	II	III	IV		I	II	III	IV	V		A	B	C	D
Stature	1.	*91*[a]	*44*	09	06	1.	*91*	*35*	11	08	03	1.	*76*	*45*	−08	−04
Sitting Height	2.	*98*	−11	15	06	2.	*96*	04	15	06	01	2.	*80*	01	00	−03
Shoulder Breadth	3.	11	24	*57*	12	3.	16	14	*53*	09	25	3.	03	19	*49*	05
Hip Breadth	4.	09	28	*76*	−02	4.	12	27	*67*	01	03	4.	01	22	*64*	−02
Span	5.	*47*	*75*	17	15	5.	*60*	*54*	22	14	*44*	5.	*52*	*55*	15	02
Chest Breadth	6.	10	14	*84*	14	6.	11	15	*73*	14	−01	6.	01	10	*65*	10
Chest Depth	7.	01	−01	*67*	05	7.	00	03	*58*	−03	−04	7.	−07	−09	*59*	−03
Head Length	8.	13	05	*32*	*64*	8.	13	06	*29*	*58*	−02	8.	09	00	*23*	*51*
Head Breadth	9.	−06	10	00	*81*	9.	−04	08	02	*64*	08	9.	−08	11	05	*61*
Head Height	10.	08	−01	−08	*70*	10.	08	−01	−06	*56*	02	10.	10	01	−11	*57*
Hand Length	11.	*37*	*81*	13	20	11.	*44*	*73*	15	22	06	11.	*46*	*71*	07	07
Hand Breadth	12.	00	*78*	*33*	−11	12.	08	*76*	*31*	−02	01	12.	01	*67*	*34*	−07

[a] Decimal points omitted in all matrices.

tests with zero specific variance may indicate only that some factors in the domain are not represented at all in the battery of tests. Zero specific variance leads logically to the use of squared multiple correlations as communalities.

Specific vectors were defined as the n orthogonal vectors obtained from iterating upon the vector sums of corresponding unit test vectors and unit unique vectors. These unit orthogonal vectors have the salient features of theoretically defined specific vectors: They

1. are orthogonal,
2. have high correlations with the corresponding test and unique vectors,
3. have low correlations with the non-corresponding test and unique vectors,
4. are identical with the unique vectors when the unique vectors are orthogonal.

The squared correlation of the i-th unique vector with the i-th specific vector is the proportion of the variance of the unique vector accounted for by the specific vector. Multiplying this squared correlation by the squared correlation of the unique vector with the i-th test vector weights the maximum uniqueness value by the proportion of the unique variance accounted for by factors fractionally and poorly represented in the i-th test. The weight matrix D_j can then be applied to the regression equation for predicting one test from the remaining test to obtain

$$F = C + D_j^2D^2F'^{-1}.$$

The weighted regression equation then states that the error is $D_j^2D^2$ rather than D^2. As a matrix of uniqueness values the elements of $D_j^2D^2$ have the following properties not shared by the Guttman-Harris uniqueness values:

1. they are always positive,
2. they are always less than the elements of D^2 given the restrictions on the correlations specified earlier and they probably are always greater than "true" values.

Weighting the correlation matrix by $D_j^{-1}D^{-1}$, analogous to communalities, leads to the formulation of the common vectors as the row vectors of

$$D_jDE(P^{1/2} - P^{-1/2}).$$

$E(P^{1/2} - P^{-1/2})E'$ is congruent to diag. $(I_p, -I_{n-p})$ with index integer p. The index integer p is the same as Guttman's strongest lower bound for the number of factors in the population when $D^{-1}RD^{-1}$ is analyzed instead of $D_i^{-1}D^{-1}RD_i^{-1}D^{-1}$. Here, however, the index integer may vary with the uniqueness values; its function is to indicate the maximum number of sample factors to be obtained from a given matrix of uniqueness values, weighted or unweighted. The index integer corresponding to $D_i^2D^2$ will be equal to or greater than the index integer corresponding to D^2. However, the use of $D_i^2D^2$ will generally result in fewer final factors than the use of D^2 and should result in a simple structure better than or as good as that obtained from the use of D^2, provided a simple structure exists in the data. Comparison of values of $(P^{1/2} - P^{-1/2})$ and $(P - I)^{1/2}$ shows that $(P_i - 1)^{1/2} > P_i^{1/2} - P_i^{-1/2}$. Furthermore,

$$D_i\ DE(P - I)^{1/2}$$

is a matrix for the reduced correlation matrix with communalities in the diagonals whereas $D_iDE(P^{1/2} - P^{-1/2})$ is only a matrix of predicted scores.

Since, the degrees of freedom available for testing p factors, p being the index integer before described, are

$$d.f. = [r^2 - (n + p)]/2$$

where $r = n - p$, the inequality $r^2 > n + p$ must hold if p factors can be tested for significance. For small correlation matrices it may be that

$$r^2 < n + p.$$

Thus, the maximum number of factors for a given $D_i^2D^2$ or D^2 might be too many factors to test for significance.

The matrix

$$EP(P - I)^{-1}E' = EP^{1/2}(P^{1/2} - P^{-1/2})^{-1}E'$$

should have small inner products as elements but will have large inner products when the latent roots P_i are in the range between 2 and 1. Therefore, it seems that the number of factors to be extracted should not be more than the number of roots, P_i, equal to or greater than 2. If attention is centered upon common factors determined solely by the side elements of the correlation matrix, the number of latent roots of the correlation matrix above unity must

constitute its maximum number of factors. If, upon rotation, one essentially null factor is found, factors of $D_jDE(P - I)^{1/2}$ should be taken from the factor matrix, one at a time, until only one null factor appears. This procedure corresponds to Thurstone's recommendation that one more factor than necessary be taken before rotation, with one factor to be discarded after rotation if uninterpretable or null. The idea is that simple structure is the final arbiter of the number of usable factors. The procedure of taking factors described will result in cleaner simple structures than the practice of taking many more factors for rotation than are finally retained provided: a simple structure exists in the data.

The process of obtaining final factors should result in factor structures and in a number of factors similar to those obtained from an analysis in the manner of Thurstone.

REFERENCES

Butler, John M., Rice, L. N., and Wagstaff, A. *Quantitative Naturalistic Research.* N. Y.: Prentice Hall, 1963.

Guttman, Louis. Multiple Linear Prediction and the Resolution into Components. *Psychometrika,* 1940, 5, 75–99.

Guttman, Louis. Image Theory for the Structure of Quantitative Variates. *Psychometrika,* 1953, 18, 277–296.

Guttman, Louis. "Best Possible" Systematic Estimates of Communalities, *Psychometrika,* 1956, 21, 273–286.

Harris, Chester W. Some Guttman-Rao Relationships. *Psychometrika,* 1962, 27, 247–263.

Harris, Chester W. Canonical Factor Models for the Description of Change. In Harris, C. W. (Ed.), *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1963.

Kaiser, Henry. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika,* 1958, 23, 187–200.

Kaiser, Henry. Image Analysis. In Harris, C. W. (Ed.) *Problems in Measuring Change.* Madison: University of Wisconsin Press, 1963.

Lawley, D. N. The Estimation of Factor Loadings by the Method of Maximum Likelihood. *Proceedings of the Royal Society of Edinburgh,* 1940, 60, 64–82.

Rao, C. R. Estimation and Tests of Significance in Factor Analysis. *Psychometrika,* 1955, 20, 93–111.

Thurstone, L. L. Factor Analysis of Body Types. *Psychometrika,* 1946, 11, 15–21.

Thurstone, L. L. *Multiple Factor Analysis.* Chicago: University of Chicago Press, 1947.

Educational and Psychological Measurement
1966, 26, 565-576.

TENSION IN FRESHMEN AND SENIOR ENGINEERING STUDENTS[1]

BEN H. ROMINE, JR.

AND

W. SCOTT GEHMAN
Duke University

According to the Taylor-Spence (1952) hypothesis, anxiety is said to facilitate performance when the performance task is a simple one and to inhibit performance when the performance task is complex. In accordance with this hypothesis, Katzenmeyer (1958) and Spielberger and Katzenmeyer (1959) have concluded that anxiety tends to interfere with academic performance, and Sarason (1957) found significant negative correlations between test anxiety and academic grade average. On the other hand, Sarason (1957) also found significant positive correlations between general anxiety and academic grade average, and Suinn (1964) found that anxiety level as measured by the Taylor Manifest Anxiety Scale (Taylor, 1953) is less predictive of academic performance than anxiety-coping ability.

If a conclusion can be drawn from the previously cited studies, it is that much more research is needed before any definite statements can be made about anxiety and its effects on academic performance.

Sanford (1962) has been concerned with anxiety in college students and has concluded that the root of anxiety among college students is found in a conflict between impulse, conscience, and ego. According to Sanford's (1962) hypothesis, this conflict typically manifests itself among college students in the authoritarian per-

---

[1] All computations for this research were made through the courtesy of the Duke University Computing Laboratory.

sonality. Other investigators, some of which are: Webster, 1958; Freedman, 1961; Webster, Freedman, and Heist, 1962; and Lehman, 1963, have concluded that freshman college students tend to evidence more authoritarian personality traits than do senior college students. The authors interpret the results of these studies as suggesting that college freshmen would tend to report more anxiety or tension than college seniors.

The authors feel that the term "anxiety" is not adaptable to the concepts dealt with in the present study. English and English (1958) offer six definitions of "anxiety." Without exception, these definitions characterize anxiety by reference to such terms as "fear," "fearfulness," "unpleasantness," or "avoidance." Olds (1955) concluded that a state of tension may be characterized by either positive (approach) or negative (avoidance and/or withdrawal) responses. Therefore it seems feasible to allow for the possibility of responses other than "fear," "unpleasantness," and/or "avoidance" responses to situations assumed to be tension-producing.

English, *et al.* (1958) define "tension" as "a condition of the organism marked by restless activity, by pressure to act and readiness to act (but with no necessary implication of directed action)." (English, *et al.*, 1958, p. 546.) In the present context, the authors feel that the term "tension" has the advantage of including the possibility of both positive (approach) and negative (avoidance and/or withdrawal) responses. For this reason, the term "tension" is employed in this study in lieu of the term "anxiety."

The purpose of the study is to determine whether or not there are any differences in the degree and kind of tension as reported by groups of freshmen and senior college students. Tension is defined in operational terms as that personality variable measured by Endler, Hunt, and Rosenstein's Stimulus-Response Inventory of Anxiousness (SRI).

It is hypothesized that:

1. The mean total tension score earned by the freshmen subjects will be higher than and statistically significantly different from the mean total tension score earned by the senior subjects. (Degree of tension.)
2. On each of the 14 mode-of-response scales, the freshmen sub-

jects will earn a mean score that is higher than and statistically significantly different from the mean score earned by the senior subjects. (Kind of tension.)

3. On each of the 11 situational scales, the freshmen subjects will earn a mean score that is higher than and statistically significantly different from the mean score earned by the senior subjects. (Kind of tension.)

Procedures

Subjects. The subjects were 86 freshmen male engineering students and 34 senior male engineering students enrolled in the Duke University College of Engineering for the academic year of 1964–1965. The freshmen were in their second semester of residence and the seniors were in the second semester of their fourth year of residence. The freshmen were distributed in age from 17–19; the seniors were distributed in age from 20–24.

Method. The SRI (Endler, *et al.*, 1962) was administered in accordance with standardized instructions to both freshmen and seniors the week following their return from spring vacation. The inventory was introduced to the subjects as a means of studying people's reactions to and attitudes toward various types of situations. The subjects were assured that their responses would be held in strict confidence.

The freshmen subjects completed the inventory in four groups. One of the authors administered the inventory to all freshmen groups. The senior subjects completed the inventory in two groups. A professor of engineering administered the inventory to 14 seniors who were in his class. One of the authors administered the inventory to the remaining 20 seniors. As an incentive to and a reward for participation, the researchers offered all subjects an interpretation of their scores as compared with scores earned by all participants.

Analysis of the data. In order to determine the degree of tension, the mean score for both freshmen and seniors, an F test of homogeneity, and a t test of the significance of the difference between means were computed for all responses to the inventory. The mean score represented a total tension score.

In order to investigate the kinds of tension reported by freshmen and seniors, mean scores earned by both groups and t tests of

the significance of the difference between means were computed for each of the 14 mode-of-response scales and for each of the 11 situational scales.

Certain of the mode-of-response and situational scale scores were grouped into three mode-of-response factors and three situational factors following the factor analysis of Endler, *et al.* (1962). Mean scores earned by both groups and *t* tests of significance of the difference between means were computed for each of the factors.

The situational scales which, in the opinion of the authors, represented activities commonly associated with the college experience were combined, i.e., "Going into a psychological experiment," "Going to meet a new date," "Getting up to give a speech before a large group," "Going to a counseling bureau to seek help in solving a personal problem," "Going into an interview for a very important job," and "Entering a final examination in an important course." The mean scores earned by both groups and a *t* test of significance of the differences between means were computed for this artifactual variable. The situational scales which, in the opinion of the authors, represented activities not commonly associated with the college experience were combined, i.e., "Just starting off on a long automobile trip," "Crawling along a ledge on a high mountain side," "Starting out in a sail boat onto a rough sea," "Entering a competitive contest before spectators," and "Alone in the woods at night." The mean scores earned by both groups and a *t* test of the significance of the difference between means were computed for this artifactual variable.

All *t* ratios reported in this study were evaluated for level of confidence by means of a table of critical values of *t*, using the .05 and .01 levels of significance for a one-tailed test.

Results and Discussion

Table 1 presents the mean scores and the *t* values of scores earned by freshmen and senior engineering students on the three mode-of-response factors, the 14 mode-of-response scales, and the total tension score.

It can be seen from the data presented in Table 1 that the *t* test of the significance of the difference between means computed for the total tension score revealed no statistically significant difference between the mean total tension scores earned by the freshmen and

TABLE 1

Mean Scores and t Values of Scores Earned by Freshmen and Senior Engineering Students on the Three Mode-of-Response Factors, the Fourteen Mode-of-Response Scales and the Total Tension Score

Mode-of-Response Factor Mode-of-Response Scale	Fr. $\bar{X}$ ($N = 86$)	Sr. $\bar{X}$ ($N = 34$)	t
Exhilaration, enjoyment, approach	2.9197	2.9599	0.4167
Feel exhilarated and thrilled	3.0941	3.1150	0.1918
Enjoy the challenge	2.7082	2.7995	0.8270
Seek experiences like this	2.9567	2.9652	0.0732
Distress, disruption, avoidance	2.2336	2.1807	0.5879
Heart beats faster	2.9894	2.9438	0.3616
Get an uneasy feeling	2.7875	2.6497	1.0732
Emotions disrupt actions	1.7611	1.7727	0.1004
Want to avoid situation	2.3055	2.2487	0.5241
Become immobilized	1.3245	1.2888	0.4426
Autonomic reactions	1.6600	1.5134	1.8876*
Need to urinate frequently	1.5359	1.4652	0.5810
Experience nausea	1.2495	1.1390	1.4511
Get full feeling in stomach	1.5159	1.4064	0.8816
Mouth gets dry	1.9873	1.7086	1.9060*
Perspire	2.4852	2.1979	1.9675*
Have loose bowels	1.1860	1.1631	0.2791
Total Tension Score	2.1450	2.0720	1.1121

* Significant at or beyond the .05 level of confidence.

senior samples. Therefore, the hypothesis that the mean total tension score earned by freshmen subjects will be higher than and statistically different from the mean total tension score earned by the senior subjects was not supported. The F ratio resulting from a comparison of the total tension scores was 1.2369 with one degree of freedom.

The t test of the significance of the difference between means computed for the various mode-of-response scales revealed that freshmen earned mean scores that were significantly higher than the mean scores earned by seniors on the "perspire" and "mouth gets dry" scales. Therefore, the hypothesis that on each of the 14 mode-of-response scales the freshmen subjects will earn a mean score that is higher than and statistically significantly different from the mean score earned by the senior subjects is partially supported.

The t test computed for the various mode-of-response factors revealed that freshmen earned a mean score that was statistically

significantly higher than the mean score earned by seniors on the "autonomic reactions" factor scale.

Table 2 presents the mean scores and t values of scores earned by freshmen and senior engineering students on the three situational factors and the 11 situational scales.

TABLE 2

Mean Scores and t Values of Scores Earned by Freshmen and Senior Engineering Students on the Three Situational Factors and the Eleven Situational Scales

Situational Factors / Situational Scales	Fr. $\bar{X}$ ($N = 86$)	Sr. $\bar{X}$ ($N = 34$)	t
Ambiguous situations	1.6836	1.7742	1.6194
Auto trip	1.5914	1.5840	0.1224
Psychological experiment	1.7757	1.9643	2.5433**
Inanimate threats	2.0628	2.0420	0.2351
Ledge on high mountain side	2.5565	2.4727	0.6362
Sail boat on rough sea	1.7135	1.7290	0.1498
Alone in woods at night	1.9186	1.9244	0.0583
Interpersonal-status-threatened	2.3211	2.1674	1.9639*
New date	1.9493	1.7668	2.1963*
Speech before large group	2.7101	2.4916	1.8077*
Counseling bureau	2.3331	2.2164	1.1929
Competitive contest	2.1055	2.0924	0.1219
Interview	2.2558	1.9076	3.5277**
Final exam	2.5731	2.5294	0.3714

* Significant at or beyond the .05 level of confidence.
** Significant at or beyond the .01 level of confidence.

It can be seen from the data presented in Table 2 that freshmen earned scores that were significantly higher than the scores earned by seniors on the following scales: "Going to meet a new date," "Getting up to give a speech before a large group," and "Going into an interview for an important job." Therefore, the hypothesis that on each of the 11 situational scales the freshmen subjects will earn a mean score that is higher than and statistically significantly different from the mean score earned by the senior subjects was partially supported.

On the "Going into a psychological experiment" scale, the seniors earned a mean score which was significantly higher than the mean score earned by the freshmen. The implication of this reversal of the tendency for freshmen to report more tension than seniors is not clear. However, it is the opinion of the authors that since the seniors have had at least one course in psychology, they

might be more sophisticated with regard to experiments involving deception, noxious stimuli, and the like.

The *t* test of the significance of the difference between mean scores for the various situational factors revealed that freshmen earned a mean score that was significantly higher than the mean score earned by seniors on the "interpersonal status threatened factor" scale.

Upon further study of the data, the authors recognized the emergence of several patterns of responses to both the mode-of-response and the situational scales which seemed worthy of further analysis. For the two groups, six scales yielded absolute mean differences which, according to the null hypothesis, would be expected to occur by chance alone on only one out of four administrations of the instrument (.25 level of confidence). With 118 degrees of freedom, a *t* value of 0.6766 is required to establish the .25 level of confidence. The authors are aware of the dangers inherent in employing levels of confidence less than .05 in identifying trends and tendencies. Therefore, in the discussion to follow, the reader must bear in mind the tentativeness of the suggestions and avoid reaching conclusions in the absence of statistical support.

As can be seen from the data presented in Table 1, of the mode-of-response scales which yielded *t* values of 0.6766 (.25 level of confidence) or greater, but did not meet an acceptable level of significance adopted for this study, the freshmen earned higher mean scores than the seniors on the following: "Get an uneasy feeling," "Get full feeling in stomach," and "Experience nausea." The "Get an uneasy feeling" scale was included in the "distress, disruption, and avoidance" factor in the original factor analysis (Endler, *et al.*, 1962). Endler, *et al.* (1962) state that this scale "point(s) to autonomic reactions." (Endler, *et al.*, 1962, p. 27.) The "Get full feeling in stomach" and "Experience nausea" scales were included in the "autonomic reaction" factor in the original factor analysis (Endler, *et al.*, 1962). With regard to the six scales included in the "autonomic reactions" factor, it can be noted that in every case the freshmen earned higher mean scores than the seniors, although not always statistically significantly higher.

As noted previously, the freshmen earned a mean score that was statistically significantly higher than the seniors on the "autonomic reactions factor" scale. It is the authors' opinion that this finding

suggests a tendency for freshmen college students to respond to tension-producing situations of the nature of those presented in the S-R Inventory (Endler, *et al.*, 1962) with a greater degree of autonomic reactions than do senior college students. From the data at hand, the implication of this finding is a matter for conjecture. One possible interpretation is that college freshmen, having met and dealt with fewer tension-producing situations than college seniors, have not developed the tension-reducing mechanisms which are available to the more experienced seniors. Stated in dynamic terms, Sanford (1962) has hypothesized that in the case of the college freshmen

> . . . the controls developed for the purpose of inhibiting impulses are still unseasoned and uncertain; they are likely to operate in a rigid manner, that is, to be rather overdone, as if the danger of their giving way altogether were still very real. (Sanford, 1962, p. 260.)

If the Sanford (1962) hypothesis is accepted, it may be said that the college freshman, whose impulse-control mechanisms are "still unseasoned and uncertain" (Sanford, 1962, p. 260), would tend to report more autonomic reactions to tension-producing situations than would college seniors who have had more opportunities to learn and employ more highly-developed control mechanisms.

As can be seen from the data presented in Table 1, of the mode-of-response scales which yielded t values of 0.6766 (.25 level of confidence) or greater, but did not meet acceptable levels of confidence adopted for this study, the seniors earned a mean score which was higher than that earned by freshmen on the "Enjoy the challenge" scale which was included in the "exhilaration, enjoyment, and approach" factor in the original factor analysis (Endler, *et al.*, 1962). It is noteworthy that on all three of the scales included in the "exhilaration, enjoyment, and approach" factor, the seniors earned higher mean scores than the freshmen although none of the differences were statistically significant.

As can be seen from the data presented in Table 2, of the situational scales which yielded t values of 0.6766 (.25 level of confidence) or greater, but did not meet the level of confidence adopted for this study, the freshmen earned a mean score which was higher than that earned by the seniors on the "Going to a counseling

bureau to seek help in solving a personal problem" scale. This scale was included in the "interpersonal status threatened" factor of the original factor analysis (Endler, *et al.*, 1962). With regard to the six scales included in the "interpersonal status threatened" factor, it can be noted that in every case the freshmen earned a higher mean score than the seniors, although not always statistically significantly higher. As has been previously noted, the freshmen earned a mean score that was significantly higher than the seniors on the "interpersonal status threatened" scale.

The implication of this finding is open for conjecture. One possible explanation is that college freshmen, having had fewer opportunities (relative to college seniors) to meet and deal with situations in which their interpersonal status is threatened, would tend to report more tension than seniors. Vogel, Raymond, and Lazarus (1963) concluded that subjects with past successes in meeting achievement and affiliative needs experienced less stress (tension) in tension-producing situations than subjects with a history of failure in these areas.

In addition to the mode-of-response and situational scales which yielded t values of 0.6766 (.25 level of confidence) or greater, but did not meet acceptable levels of confidence, a comparison of the total tension mean score earned by the freshmen and senior groups also yielded a t value greater than 0.6766 (.25 level of confidence). The freshmen tended to earn a higher total tension mean score than did the seniors.

It is hypothesized by the authors that had the S-R Inventory (Endler, *et al.*, 1962) been administered earlier in the academic year, more significant differences between mean scores earned by freshmen and seniors would have been revealed. The basis for mentioning this possibility lies partially in the fact that the trends identified in the present study seem to suggest some significant differences between mean scores earned by freshmen and seniors on certain of the mode-of-response and situational scales and factors. Much more important than this, it would seem, is the report by Freedman (1961) that "studies of juniors, sophomores, and freshmen in their second semester indicate that . . . senior-freshmen changes are not linear from year to year but tend to take place quite early." (Freedman, 1961, p. 22.)

After analyzing the data and observing the tendency for fresh-

men to earn higher mean scores than seniors on those scales which might intuitively be considered to involve activities commonly associated with the college experience, the authors combined all such scale scores and computed a t test of significance of the difference between the means of the freshmen and senior groups.

Table 3 presents the mean scores and t values of scores earned by freshmen and senior engineering students on the combined situational scales which involve activities not commonly associated with the college experience.

TABLE 3

Mean Scores and t Values of Scores Earned by Freshmen and Senior Engineering Students on the Combined Situational Scales Which Involve Activities Commonly Associated with the College Experience, and the Combined Situational Scales Which Involve Activities Not Commonly Associated with the College Experience

Combined Scale	Fr. $\bar{X}$ (86)	Sr. $\bar{X}$ (34)	t
College experience	2.2662	2.1460	1.6935*
Non-College experiences	1.9771	1.9605	0.2282

* Significant at or beyond the .05 level of confidence.

As can be seen from the data presented in Table 3, the t test of significance of differences between means computed for the combined situational scales taken to represent situations which involve activities commonly associated with the college experience, revealed that the freshmen earned a mean score which was higher than, and statistically significantly different from, the mean score earned by seniors. The t test of significance of the difference between means computed for the combined situational scales which involved activities not commonly associated with the college experience, revealed no differences between the mean scores earned by freshmen and seniors.

The authors interpret this observation as suggesting that as college students gain experience, the tension likely to be produced by situations such as those presented in the S-R Inventory (Endler, *et al.*, 1962) tends to be diluted.

Summary

This study represents an attempt to determine whether or not there are any differences in the degree and kind of tension as re-

ported by groups of freshmen and senior engineering students.

In operational terms, tension is defined as that personality variable measured by the Stimulus-Response Inventory of Anxiousness (SRI) as constructed by Endler, Hunt, and Rosenstein (Endler, *et al.*, 1962).

Eighty-six freshmen and 34 senior male engineering students completed the S-R Inventory under standardized procedures.

In order to determine the degree of total tension reported by freshmen and seniors, the mean score for both groups and a *t* test of significance of differences between means were computed for all responses to the inventory. In order to investigate the kinds of tension reported by freshmen and seniors, mean scores earned by both groups and *t* tests of significance were computed for each of the 11 situational scales and for each of the 14 mode-of-response scales. In addition to these scores, certain of the mode-of-response and situational scale scores were grouped into three mode-of-response and three situational factors following the factor analysis of Endler, *et al.* (1962). Mean scores earned by both groups and *t* tests of the significance of differences between means were computed for each of these factors. The situational scales which, in the opinion of the authors, represented activities commonly associated with the college experience were combined. The mean scores earned by both groups and a *t* test of the significance of the difference between means were computed for this artifactual variable. The situational scales which, in the opinion of the authors, represented activities not commonly associated with the college experience were combined. The mean scores earned by both groups and a *t* test of the significance of the difference between means were computed for this artifactual variable.

An analysis of the data revealed no statistically significant difference between freshmen and seniors with regard to the degree of total tension as defined by the mean scores earned on the S-R Inventory (Endler, *et al.*, 1962). There were statistically significant differences between the mean scores for freshmen and seniors with regard to the kind of tension as measured by certain mode-of-response scales presented in the inventory. Furthermore, there were statistically significant differences between the mean scores for freshmen and seniors with regard to the types of situations which produced tension.

The authors recognize the S-R Inventory (Endler, *et al.*, 1962) as a valuable research instrument.

REFERENCES

Endler, N. S., Hunt, J. McV., and Rosenstein, A. J. An S-R Inventory of Anxiousness. *Psychological Monographs*, 1962, No. 17 (Whole No. 536).

English, H. B. and English, Ava C. *A Comprehensive Dictionary of Psychological and Psychoanalytical Terms*. New York: David McKay Co., 1958.

Freedman, M. B. Influence of College Experience on Personality Development. *Psychological Review*, 1961, 8, 21–22.

Katzenmeyer, W. G. A Study of the Relationship between a Personality Variable and Academic Performance. Unpublished master's thesis, Duke University, 1958.

Lehman, I. J. Changes in Critical Thinking, Attitudes and Values from Freshmen to Senior Years. *Journal of Educational Psychology*, 1963, 54, 305-315.

Olds, J. Physiological Mechanism of Reward. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation: 1955*. Lincoln, Nebraska: University of Nebraska Press, 1955.

Sanford, N. The Developmental Status of the Entering Freshman. In N. Sanford (Ed.) *The American College*. New York: John Wiley & Sons, Inc., 1962, Pp. 314–353.

Sarason, I. Test Anxiety, General Anxiety and Academic Performance. *Journal of Consulting Psychology*, 1957, 21, 485–490.

Spielberger, C. D. and Katzenmeyer, W. G. Manifest Anxiety, Intelligence, and College Grades. *Journal of Consulting Psychology*, 1959, 23, 278–287.

Suinn, R. M. A Factor Modifying the Concept of Anxiety as an Interfering Drive. *Journal of General Psychology*, 1964, 70, 134–140.

Taylor, Janet A. A Personality Scale of Manifest Anxiety. *Journal of Abnormal and Social Psychology*, 1953, 48, 285–290.

Taylor, Janet A. and Spence, K. W. The Relationship of Anxiety to Performance in Serial Learning. *Journal of Experimental Psychology*, 1952, 44, 61–64.

Vogel, S., Raymond, F., and Lazarus, P. Intrinsic Motivation and the Psychology of Stress. In Martha T. Mednick & S. A. Mednick (Eds.), *Research in Personality*. New York: Holt, Rinehart & Winston, Inc., 1963, Pp. 280–314.

Webster, H. Changes in Attitude During College. *Journal of Educational Psychology*, 1958, 49, 109–117.

Webster, H., Freedman, M. B., and Heist, P. Personality Changes in College Students. In N. Sanford (Ed.), *The American College*. New York: John Wiley & Sons, Inc., 1962, Pp. 811–846.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 577-582.

IMPROVED HIERARCHICAL SYNDROME ANALYSIS OF DISCRETE AND CONTINUOUS DATA

LOUIS L. McQUITTY
Michigan State University

IN the isolation of statistically defined types, Hierarchical Syndrome Analysis has the advantage of being simple, brief, and applicable to both discrete and continuous data. In its original development, it does, however, have shortcomings; once Individual i has classified with Individual j then he cannot by himself classify with any individual other than j, and this is true even though j and k each have i most like them. In order to correct for this limitation, a version was developed which classifies i with each j and k in the above circumstance, but the latter version has the disadvantage of being cumbersome and of yielding unnecessarily complicated classifications (McQuitty, 1960).

An Improvement

An improvement is realized by combining certain features of each (a) Hierarchical Analysis by Reciprocal Pairs (McQuitty, 1964, 1966), and (b) Hierarchial Syndrome Analysis (McQuitty, 1960); the theory of types from (a) is applied to a new version of the approaches of (b). More specifically, the problem between i and j, as outlined above, is solved by defining types in such a fashion that the problem disappears; the analysis then proceeds without internal inconsistencies.

An hierarchical classification is a successive categorization of *imperfect* types from one level into more exacting representation of types at the next higher level. Each successive level contains fewer categories than the previous one. Each category reflects only one *imperfect* type; it is defined by the characteristics of the individ-

uals in the category. Single individuals represent *imperfect* types at the first level of classification.

An *hierarchical* type is represented by a pair of *imperfect* types from the next lower level of classification; each of these *imperfect* types is more like the other one of them than it is like any other *imperfect* type at the same stage of classification.

Hierarchical types are intermediate between those represented in reality by single individuals and the *pure* types of theory, which are approached but never quite realized, as individuals are classified in higher and higher levels.

By the above definition, ij constitutes a type if, and only if, the association between i and j is larger than that between i and k, and on the other hand, ik constitutes a type if, and only if, the association is reversed; i and k is larger than i and j. If the associations are tied, neither ij nor ik constitute a type. This is the *principle of classification by reciprocal pairs* (McQuitty, 1964, 1966).

For the purpose of further clarification assume:

a. i is highest in Column j

b. i is highest in Column k

c. j is highest in Column i; k is not highest in Column i; i and j are reciprocal

d. $r_{ij} > r_{ik}$

Then by the above definition of types, i and j would join to form an *hierarchical* type which is a better representative of a *pure* type than is either i or j separately. In the analysis, the two poor representatives, i and j, of a single *hierarchical* type, ij, would be replaced by ij. It is now irrelevant that i is highest in Column k, and the problem has been eliminated.

The above discussion illustrates a very important characteristic of the method. The method is designed to improve and purify the data as it proceeds; each successive category is intended to become a better representative of a type as it exists in theory.

In summary, the method searches for evidence of types. A reciprocal pair (i highest with j and j highest with i) is accepted as such evidence. The members of a pair are combined and the two jointly are accepted as a better representative of a type than is either separately.

The above definition of an *hierarchical* type enables us to solve

another problem central to our classification procedures. It justifies the *classification assumption* (McQuitty, 1960): an *hierarchical* type has as much association with another *hierarchical* type as the smallest association represented by all pairs of persons involved in either one or both of the two types. This assumption maintains that the upper limit in the size of the association (as represented by the lowest pair) is, in fact, also the lower limit.

In order to perform perfectly, the assumption need not be perfectly valid; it need only be relatively valid. It must preclude non-typal categories from competing successfully with typal categories, and non-typal categories are characterized by extremes in low pairs, while typal categories are in general just the opposite.

Method

The method is illustrated by application to the data of Matrix 1, Table 1, coefficients of correlations between people for the second through the sixth person of a matrix from Stephenson (1953, p. 169).

TABLE 1

An Analysis of Coefficients of Correlation between People

	B	C	D	E	F
B		60	49	24	43
C	60		62	46	51
D	49	62		40	57
E	24	46	40		56
F	43	51	57	56	

Matrix 1

	B	CD	E	F
B		49	24	43
CD	49		40	51
E	24	40		56
F	43	51	56	

Matrix 2

	B	CD	EF
B		49	24
CD	49		40
EF	24	40	

Matrix 3

	BCD	EF
BCD		24
EF	24	

Matrix 4

Note.—Data of Matrix from Stephenson (1953, p. 169); the entries have been rounded from three to two place numbers. The decimal points have been omitted.

The first step is to underline the highest entry of each column of Matrix 1, 60 for Column B, 62 for C, 62 for D, 56 for E, and 56 for F. Then, by selecting the highest underlined entry, one selects the highest entry in the matrix; it is 62 and appears in both Columns C and D.

Matrix 2 is formed from Matrix 1 by collapsing Columns C and D of Matrix 1 into a single Column CD in Matrix 2. In determining the entry for the top cell of Column CD (Row B—Column CD), one looks to two cells of Matrix 1: (a) Row B—Column C with an entry of 60, and (b) Row B—Column D with an entry of 49. The

smaller of these two entries, 49, is the entry of Row B—Column CD, Matrix 2. In case of a tie between these two values, the tied value is used.

The other entries of Column CD, Matrix 2, are determined in a similar fashion. The entries of Row CD are the same as those of Column CD. The other entries of Matrix 2 are realized by copying the corresponding entries from Matrix 1.

Each subsequent matrix is formed from its predecessor in the same fashion as Matrix 2 was formed from Matrix 1. For example, the highest entry in Matrix 3 is 49 in Columns B and CD. In forming Matrix 4, these two columns are collapsed into Column BCD. The entry for Row EF—Column BCD of Matrix 4 is the smaller of the following two entries of Matrix 3: (a) Row EF—Column B which is 24, and (b) Row EF—Column CD which is 40. Consequently, 24 is the entry for Row EF—Column BCD (and for Row BCD—Column EF) of Matrix 4.

Results

Matrices 2 through 4 reflect the classification. It is summarized in Figure 1 which shows that Persons C and D first joined one another with an index of 62, and were followed in order by E and F with an index of 56, B and CD with an index of 49, and finally BCD and EF with an index of 24.

A Critique of the Method

In the present approach, we eliminated but one reciprocal pair at a time (the highest one) in forming a new matrix. We could have eliminated all of them each time we formed a new matrix, and this is the way in which the *replacement version,* as originally described, does in fact proceed (McQuitty, 1960). The *highest-pair version,* as here described is, however, less complicated.

A word of precaution is essential. The highest pair of a matrix is always reciprocal, and if an analysis yields only one reciprocal pair in each of its successive matrices, it is well to realize that this could occur from chance alone, but at the same time, does not need to be due to chance alone. One needs additional evidence, such as the meaningfulness of the categories, to support an interpretation other than chance.

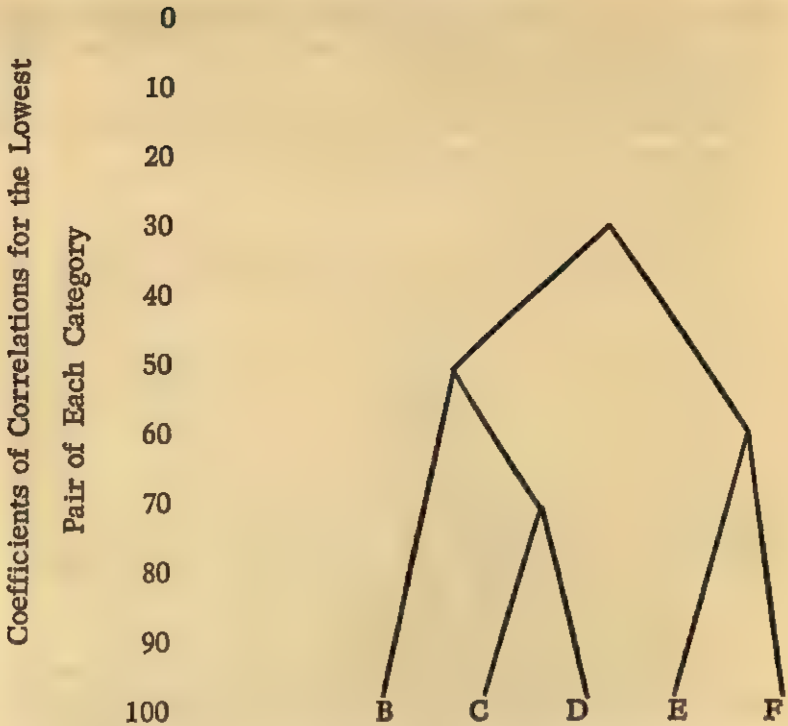


Figure 1. Classification of people from a matrix of intercorrelations between pairs of them.

Summary

This paper develops a new version of Hierarchical Syndrome Analysis, called the *highest-pair version;* this version is simpler both to describe and apply than either of the two earlier versions, the *replacement version* and the *self-checking version.*

In developing the new version, this paper shows how both the new version and the *replacement version* can be interpreted to improve and purify the data as they proceed; evidence of types is sought and then the analysis is intended to improve on the descriptions of the types as the analysis proceeds.

All three versions are applicable to both discrete and continuous data.

REFERENCES

McQuitty, L. L. Hierarchical Syndrome Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 293–304.

McQuitty, L. L. Capabilities and Improvements of Linkage Analysis as a Clustering Method. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 441–456.

McQuitty, L. L. Single and Multiple Hierarchical Classification by Reciprocal Pairs and Rank Order Types. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 253–265.

Stephenson, W. *The Study of Behavior.* Chicago: The University of Chicago Press, 1953.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 583-603.

THE EFFECT OF VARIATIONS IN THE CUE R MATRIX UPON THE OBTAINED POLICY EQUATION OF JUDGES

ARTHUR L. DUDYCHA AND JAMES C. NAYLOR
The Ohio State University

IN several earlier articles the multiple regression model has been demonstrated to be an extremely useful methodology for describing the strategies or policies of judges with respect to their responses to a set of complex stimuli (Naylor and Wherry, 1965; Wherry and Naylor, 1966). In such studies the usual procedure is to relate each of the k characteristics of the stimuli to the responses made by the judge by means of a least squares prediction equation

$$Y_i' = b_1^*X_{1i} + b_2^*X_{2i} + \cdots + b_k^*X_{ki}$$

where

Y_i' = predicted response of a judge when presented with stimulus i

$(b_1^*, b_2^*, \cdots, b_k^*)$ = least square regression weights for the judge in standard score form for each of the k stimulus dimensions. They are obtained by computing a multiple regression equation (usually over all n stimuli) using the different stimulus cues as predictors and the judge's response as the criterion.

$(X_{1i}, X_{2i}, \cdots, X_{ki})$ = values, in standard score form, for each of the different cue dimensions for stimulus i.

Now any set of n stimulus objects presented to a judge can itself be described in terms of an R matrix which is formed by intercorrelating the different cue values over all stimuli, thus

$$R = \begin{bmatrix} r_{x_1x_1} & r_{x_1x_2} & \cdots & r_{x_1x_k} \\ r_{x_2x_1} & r_{x_2x_2} & \cdots & r_{x_2x_k} \\ r_{x_3x_1} & \cdots\cdots\cdots & & r_{x_3x_k} \\ r_{x_4x_1} & & & r_{x_4x_k} \\ \vdots & & & \vdots \\ r_{x_kx_1} & \cdots\cdots\cdots & & r_{x_kx_k} \end{bmatrix}$$

Naylor and Wherry (1964) have raised the question of the importance of this R matrix as a determiner of a judge's equation or policy. Will one obtain similar policies using different sets of stimuli which in turn have different R matrices? The question would certainly seem to be a basic one, since at best the n stimuli used in an experiment represent a truly random sample of stimuli drawn from a population size N. However, in many cases, due to various practical expediencies, our experimental stimuli may not be truly "representative" of the population to which they belong, i.e., $R_{\text{EXP.}}$ will not be representative of R_{TRUE}. For example, in some circumstances the R matrix is deliberately made an identity matrix in order to allow E to assess the "independent contribution" of each cue dimension. While one may applaud the objective, it raises the question as to whether the resulting policies are truly those that one would obtain using stimuli having an R matrix more truly representative of R_{TRUE}.[1]

The purpose of the present research study was to determine what influence distorting the R matrix from R_{TRUE} would have upon the rater policy equations obtained using the multiple regression model.

Method

Experimental Stimuli

The stimuli used in the study were job "profiles"—that is, each stimulus represented a particular job which had been given a score on each of a number of different job dimensions. Subjects were required to examine each job and give it a rating in terms of its overall "Job Desirability." The job characteristics which were used were those obtained in an orthogonal re-rotation of the Baehr

[1] Of course there is always the possibility that R_{TRUE} is indeed an identity matrix. However, the authors doubt that such will often be the case.

(1954) and Ash (1954) factor analytic studies of the SRA Employee Inventory (presently known as the SRA Attitude Survey) by Wherry (1954). The R matrices from the Baehr and Ash studies (two from Baehr's study and one from Ash's) were combined into a single R matrix by computing the weighted averages for each intercorrelation of the fourteen subtests of the SRA Attitude Survey. The resulting matrix was assumed to describe the "real" world interrelationships of the 14 subtests (job traits) of the SRA Attitude Survey and was called R_{TRUE}. In a similar manner the three respective orthogonally rotated factor structures from Wherry (1954) were combined, yielding the theoretical F matrix associated with R_{TRUE}. This final F_{TRUE} matrix contained six factors—one general and five subgeneral—plus 14 specifics as being appropriate to describe the underlying dimensions necessary to explain the job traits. The six factors were (G) General Bias or Attitude, (A) Working Conditions and Requirements (status), (B) Financial Reward, (C) Immediate Supervision, (D) Effective Management and Administration, and (E) Security and Communication (see Table 1).

Distorted R Matrices

Two related factor structures with their respective correlation matrices were also created, holding the reliabilities (r_{tt}) constant. These two factor structures and their correlation matrices were distortions of F_{TRUE} and R_{TRUE}. The first distortion (D_L) was obtained simply by modifying F_{TRUE} into a completely orthogonal or independent F matrix containing the square root of the F_{TRUE} reliabilities ($\sqrt{r_{tt}}$) in the diagonal, while the second distortion (D_H) was an F matrix obtained by systematically collapsing all of the specifics of the F_{TRUE} matrix into the appropriate six factors of this factor structure. Thus, three distinct theoretical F matrices and their associated theoretical R matrices were defined to represent three theoretical sets of stimuli (see Tables 2–4).

Profile (Stimuli) Generation

Three sets of two hundred job profiles were produced (one set from each of the previously created F matrices) using the Ohio State Correlated Score Generation Method, (Wherry, Naylor, Wherry and Fallis, 1965). Each profile under each condition presented a score for that "job" on each of the fourteen traits. The

TABLE 1

Factor Structure for Condition T

Variable	Factors						Specifics														
	G	A	B	C	D	E	1	2	3	4	5	6	7	8	9	10	11	12	13	14	r_{tt}
1	57	32	−06	02	−01	05	62													10	828
2	49	51	10	20	26	−08		60								−10					994
3	49	12	57	−11	02	−09			35												723
4	47	04	34	04	39	−04		−10		39	−17	−16					−19	−20		−10	796
5	51	01	00	−01	40	−12		−10	−10		61		−20		−20	−19	−20		−10		993
6	65	11	04	50	02	08		−09				54									993
7	75	12	10	03	28	04	−10						56	−07							996
8	67	11	−08	43	10	00							−07	57							992
9	66	19	00	15	37	−05		−08		−13		−11			53						950
10	66	11	23	10	25	20				−12						34					743
11	68	03	22	15	31	17		−09						−10	−10		54				979
12	72	23	32	−04	09	11												43			880
13	61	17	10	−06	05	08	−09	−22	−15										44		696
14	66	20	12	01	03	−04														42	669

Note.—In this table, the factor loadings are given correct to two decimal places and the decimal points are omitted.

TABLE 2

(Condition T)

Trait Theoretical Intercorrelations (Above Diagonal), Reliabilities (Diagonal) and Trait Empirical Intercorrelations (Below Diagonal)

	Trait Number[a]					
	1	2	3	4	5	6
1	100	31	30	42	41	43
2	35	100	23	28	42	41
3	24	19	100	35	34	34
4	37	26	35	100	52	45
5	40	42	25	54	100	49
6	36	26	26	31	46	100

Note.—In this and the two subsequent tables, the figures are given correct to two decimal places and the decimal points are omitted.

[a] The trait numbers in this and the two subsequent tables refer to the following trait names:

1. **Working Conditions**
2. **Pay**
3. **Fellow Employees**
4. **Supervisor-Employee Relations**
5. **Security**
6. **Opportunity for Growth and Advancement**

TABLE 3

(Condition D_L*)*

Trait Theoretical Intercorrelations (Above Diagonal), Reliabilities (Diagonal) and Trait Empirical Intercorrelations (Below Diagonal)

	Trait Number					
	1	2	3	4	5	6
1	100	00	00	00	00	00
2	−09	100	00	00	00	00
3	−04	07	100	00	00	00
4	01	−01	−02	100	00	00
5	−04	−02	03	−06	100	00
6	06	04	−03	−03	03	100

TABLE 4

(Condition D_H*)*

Trait Theoretical Intercorrelations (Above Diagonal), Reliabilities (Diagonal) and Trait Empirical Intercorrelations (Below Diagonal)

	Trait Number					
	1	2	3	4	5	6
1	100	37	61	60	63	70
2	39	100	32	27	43	46
3	60	31	100	39	81	42
4	59	26	47	100	65	50
5	56	37	76	64	100	51
6	64	42	34	44	45	100

scores ranged from 1 (very poor) to 9 (very good) with 4 through 6 considered an average range. The two hundred scores for each trait were normally distributed, and the intercorrelations of profile trait scores differed only by sampling error from the appropriate theoretical "population" correlation matrix obtained directly from the underlying F matrix used to generate the profile scores.

Thus, three different sets of experimental stimuli conditions were created: Condition T, regarded as the control condition, consisted of the profiles based on the R_{TRUE} matrix and was considered as being representative of the "real" world situation; Condition D_L, which contained profiles generated from the orthogonal F matrix yielding intercorrelations approximating zero; and Condition D_H, which consisted of profiles generated from the collapsed F_{TRUE} matrix producing relatively high intercorrelations.

Because of testing time limitations it was necessary that the actual stimuli presented to each S only include six of the fourteen original job traits [(1) Working Conditions, (2) Pay, (3) Fellow Employees, (4) Supervisor-Employee Relations, (5) Security and (6) Opportunity for Growth and Advancement]. Tables 2, 3 and 4 show the theoretical and empirical R matrices for these six job traits. The six job traits describing the profiles were selected on the basis of meaningfulness to the rater population relative to the criterion, "Job Desirability," and as being the traits most likely to be relevant to the employee's perception of the work environment.

Subjects

One hundred and fourteen male introductory psychology students at Ohio State University served as raters in this study who, in doing so, partly fulfilled their course requirements of experimental participation. They were divided into three equal groups of 38 members each. The selection of raters for a particular group was based strictly on their availability for testing rather than by any random assignments to the three respective groups.

Procedure

The three sets of profiles (one set of two hundred profiles for each condition) were typed onto a 180 foot roll of typing paper which could be inserted into a specially constructed device designed to feed the paper continually into an opaque projector. Preceding

each set of two hundred profiles were three examples. Each individual profile within a condition set was successively numbered from 1 through 200, and was not identified by any specific job title. Through the use of a black template, on which the individual job traits were identified, individual job profiles were able to be displayed sequentially from 1–200 to subjects in a partially darkened room. Each profile was displayed for approximately 12–15 seconds, the length of time being controlled by *E*. The job traits were listed in the same order for all raters. Figure 1 shows a sample profile displayed on the screen as viewed by the rater.

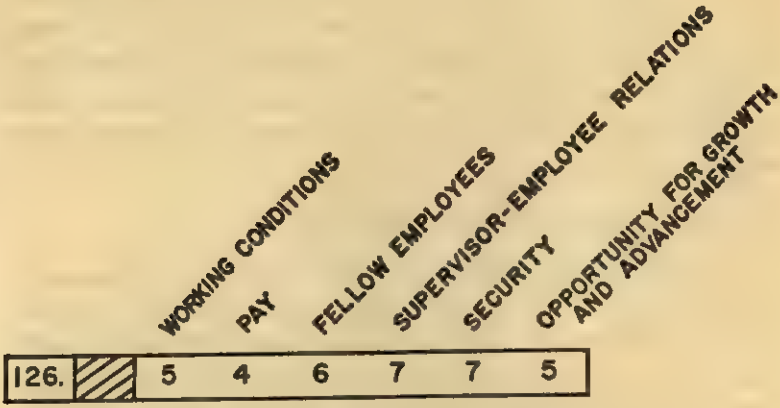


Figure 1. Sample "job" profile as viewed by the raters.

Raters were tested in groups of ten to twenty in size depending upon their availability for testing, with 38 different raters assigned to each condition. Each rater was given a set of directions describing the experiment and his task, a list of the six job traits with their descriptors, and an answer sheet with two hundred numbered blanks—one for the rater's evaluative response to each profile. Each was instructed that each individual profile was representative of a particular job. He was told to examine each profile and give it a score from 1 (very undesirable) to 9 (very desirable) in terms of the criterion, "Job Desirability."

Results

Regression or policy equations for each rater were computed using the generated profile job trait scores as predictors (X_i) and his response or profile evaluation relative to "Job Desirability" as the

criterion. These regression equations, which represent the policies or strategies used by the raters in evaluating the profiles, were then sequentially clustered within each condition using the JAN technique of Bottenberg and Christal (1961) in an attempt to identify groups of raters holding heterogeneous subpolicies. This was followed by an analysis of variance using the validity coefficients of the predictor traits in order to examine the results for possible between-cluster differences, between-condition differences, and for cluster-by-condition differences. Finally, a principal axis factor analysis followed by a varimax (Kaiser) rotation was performed on an R matrix (114 $\times$ 114) based upon the intercorrelations between the validity vector of each rater with those of all other raters across all three conditions. The factor analysis of these rater validity vectors (PROF Technique—Wherry and Naylor, 1966) was performed to compare the obtained rater factor strategies in the three conditions.

Regression Equations (Policy Predictability)

The results of the regression analyses performed upon the profile evaluations of each judge are summarized in Table 5. The table gives the squared multiple correlation coefficient (R_i^2) obtained for each rater. R_i^2 is an expression of the consistency of a rater's judgment across all two hundred profiles. The R_i^2 values show that while in general all raters exhibited a high level of intrajudge consistency, there were certainly substantial differences between raters within the respective conditions. For example, the R_i^2 values for Condition T ranged from a high of .905 (Rater 4) to a low of .522 (Rater 23); from .923 (Rater 21) to .490 (Rater 36) for Condition D_L; and from .940 (Rater 14) to .522 (Rater 8) for Condition D_H. However, there is a very distinct and systematic difference between the relative magnitudes of the R_i^2 values for each separate condition, with raters in Condition D_H being most systematic, raters in Condition D_L being least systematic, while raters in Condition T fall between the previous two conditions.

The raw score regression weights (b_j) given to each of the six predictor variables (job traits) by a rater and the associated intercept were also obtained. The data presented in Table 6 summarize the results obtained by computing the average regression weight for each of the six predictor variables across each of the

TABLE 5

Squared Multiple Correlation Coefficients (R_i^2) for the Thirty-eight Raters in Each Experimental Condition (Rater Consistency)

Rater	Condition T	Condition D_L	Condition D_H
1	859	670	915
2	874	794	875
3	774	705	912
4	905	654	914
5	827	736	912
6	807	774	910
7	877	747	920
8	879	728	522
9	847	701	913
10	891	784	911
11	701	666	827
12	872	758	829
13	804	785	890
14	833	816	940
15	872	679	923
16	891	658	894
17	858	515	907
18	765	807	900
19	858	761	909
20	831	778	896
21	846	923	900
22	852	848	916
23	522	789	869
24	836	725	887
25	875	681	868
26	862	797	834
27	800	759	889
28	868	707	910
29	857	870	840
30	847	778	912
31	883	774	859
32	848	802	901
33	823	740	879
34	897	775	896
35	896	687	904
36	853	490	896
37	882	774	915
38	899	815	907

three conditions. Note that the rank order and the relative magnitudes of the average regression weights are very similar under each of the three conditions. The job traits, Pay and Opportunity for Growth and Advancement, were quite consistently weighted high by most raters within each condition, while Fellow Employ-

TABLE 6

Average Raw Score Regression Weights ($\bar{b}_j$'s) Associated With Each of the Profile Job Traits for Each Condition

Job Traits	Conditions T	Conditions D_L	Conditions D_H	Total
Working Conditions	174 (3)[a]	154 (4)	120 (5)	149 (4)
Pay	351 (1)	461 (1)	359 (1)	390 (1)
Fellow Employees	112 (6)	132 (5)	136 (4)	127 (5)
Supervisor-Employee Relations	115 (5)	112 (6)	100 (6)	109 (6)
Security	159 (4)	199 (3)	143 (3)	167 (3)
Opportunity for Growth and Advancement	331 (2)	361 (2)	335 (2)	342 (2)

Note—In this table the figures are given correct to three decimal places with the decimal points omitted.

[a] The values in the parentheses represent the rank order of the job traits within that condition.

ees, and Supervisor-Employee Relations consistently received low weightings by most raters across conditions.

Policy Comparisons (Policy Comparability)

In order to evaluate the raters' policy equations obtained using the different underlying cue R matrices, several comparisons seemed important. First, would the same group policies emerge, that is, would the same relative weights be given to the predictor variables (stimulus cues) under each condition when averaged across all raters in that condition? Second, if there exist different subpolicies among the raters within a condition, will the *same* subpolicies emerge under each condition? The basic index of a rater's policy adopted to examine the previous questions was that rater's validity vector, since prior research by Naylor and Wherry (1964) had indicated the validity vector to be a more stable index than a vector of regression weights.

First, the raters' regression equations were sequentially clustered within each condition using JAN (Bottenberg and Christal, 1961). The JAN iterative clustering criteria minimized the differential composite squared multiple (ΔR_c^2) for the entire group, retaining optimum predictive efficiency at each stage in the clustering process. Thus, $\Delta R_c^2 = R_{c_i}^2 - R_{c_j}^2$, where $R_{c_i}^2$ was defined as the composite group predictive efficiency at stage i in the clustering process and $R_{c_j}^2$ as the composite group predictive efficiency at stage $j (j = i + 1)$, the next successive clustering stage. The composite predictive efficiency at any stage i ($R_{c_i}^2$), was defined as

$$R_{ci}^2 = \frac{SS\text{reg}_1(N_1) + SS\text{reg}_2(N_2) + \cdots + SS\text{reg}_{k-i}(N_{k-i})}{SS_{\text{total}_1} + SS_{\text{total}_2} + \cdots + SS_{\text{total}_k}}$$

where k is the original number of regression equations to be grouped and i is the number of stages (Naylor and Wherry, 1964).

Table 7 presents the computed values of R_c^2 and ΔR_c^2 for each of

TABLE 7

Composite Predictive Efficiency (R_c^2) at Each Stage of the Grouping Process for Each of the Three Conditions

	Condition T		Condition D_L		Condition D_H	
System	R_c^2	ΔR_c^2 [a]	R_c^2	ΔR_c^2	R_c^2	ΔR_c^2
38	8456	0007	7597	0002	8883	0002
37	8449	0017	7595	0026	8881	0013
36	8432	0001	7569	0061	8868	0002
35	8431	0065	7508	0002	8866	0021
34	8366	0009	7506	0009	8845	0001
33	8357	0003	7497	0013	8844	0001
32	8354	0021	7484	0003	8843	0002
31	8333	0004	7481	0047	8841	0009
30	8329	0003	7434	0045	8832	0002
29	8326	0012	7389	0017	8830	0002
28	8314	0004	7372	0007	8828	0003
27	8310	0006	7365	0081	8825	0004
26	8304	0024	7284	0007	8821	0004
25	8280	0006	7277	0031	8817	0005
24	8274	0008	7246	0009	8812	0004
23	8266	0007	7237	0031	8808	0014
22	8259	0007	7206	0016	8794	0006
21	8252	0007	7190	0015	8788	0005
20	8245	0005	7175	0012	8783	0011
19	8240	0014	7163	0010	8772	0011
18	8226	0012	7153	0056	8761	0014
17	8214	0015	7097	0014	8747	0011
16	8199	0014	7083	0072	8736	0010
15	8185	0016	7011	0050	8726	0012
14	8169	0022	6961	0017	8714	0018
13	8147	0040	6944	0042	8696	0015
12	8107	0028	6902	0021	8681	0022
11	8079	0023	6881	0025	8659	0018
10	8056	0021	6856	0036	8641	0024
9	8035	0029	6820	0056	8617	0021
8	8006	0044	6764	0073	8596	0036
7	7962	0062	6691	0168	8560	0030
6	7900	0062	6523	0072	8530	0043
5	7838	0066	6451	0092	8487	0077
4	7772	0068	6359	0098	8410	0082
3	7704	0147	6261	0087	8328	0096
2	7557	0291	6174	0228	8232	0193
1	7266		5946		8039	

Note.—In this table the figures are given correct to four decimal places and the decimal points are omitted.

[a] $\Delta R_c^2 = R_{c_i}^2 - R_{c_j}^2$, where $j = i + 1$.

the three conditions as the number of rater-clusters was sequentially and systematically reduced from 38 to a single all encompassing cluster. Table 7 shows a noticeably higher predictive efficiency (within-rater consistency) for the raters under Condition D_H as opposed to those of Condition D_L, while the Condition T raters' predictive efficiency falls somewhere between those of the other two conditions. Also, note that the between-rater consistency, which is measured by the overall drop in R_c^2 from the 38-system stage to the single-system stage or the sum of the ΔR_c^2 values, varies between conditions, where it is lowest (.084) for Condition D_H, highest (.165) for Condition D_L and between the previous conditions (.119) for Condition T.

Analysis of Variance on Policy Similarity

The data used to make the policy comparisons were obtained from the two-group clustering stage (system 2), that is, the raters' regression equations were clustered into two distinct groups within each condition. These clusters maximized the between-group difference and the within-group similarity of the regression equations. Table 8 summarized the results of an unweighted means analysis

TABLE 8

Unweighted Means Analysis of Variance Source Table

Source	df	MS	F	p
Between Subjects	113			
C (Conditions)	2	12.3270	1090.88	<.001
G (Groups)	1	.0323	2.86	<.10
C × G	2	.0015	.13	
S's/G	108	.0113		
Within Subjects	570			
T (Traits)	5	1.9500	90.70	<.001
T × C	10	.4037	18.78	<.001
T × G	5	.2075	9.65	<.001
T × C × G	10	.0273	1.27	<.25
T × S's/G	540	.0215		
Total	683			

Note.—For an unweighted-means solution, $SS_{total} \neq SS_{between\ cells} + SS_{within\ cells}$. In the ANOV model Conditions, Groups, and Traits were treated as fixed variables, Subjects as a random variable.

of variance for unequal group size performed upon the validity coefficients of the predictor traits for each rater in order to examine the results for possible between-cluster differences, between-condition differences, and for cluster-by-condition differences.

Between Profile Comparisons. The analysis revealed that the main effect of Conditions was significant at $p < .001$, but no significant Groups effect was found. Also, no significant (C $\times$ G) interaction was found.

The significant Conditions effect indicated that the policy *levels* for the three conditions were significantly different. The pattern of the vectors of average validities for the raters under each condition are shown in Figure 2, which indeed depicts the raters for the three conditions as holding somewhat similar policies but at substantially different average levels.

The nonsignificant Group effect ($F = 2.86$, $df = 1{,}108$) indicates that the policy levels for the two JAN subpolicy groups were not significantly different when averaged across all three conditions. In addition, since no significant interaction ($F = .13$, $df = 2{,}108$) was found between Conditions and Groups it would appear that similar group subpolicy levels emerged from each condition. A more comprehensive clarification of these results can be obtained by examining Figures 3, 4 and 5. From these figures it can be seen that the two subpolicies (profiles) obtained by JAN under each condition do not appear to have different average heights—thus, the nonsignificant Groups effect. Also, since the difference in profile heights or levels of the two groups does not appear to change over the three conditions the nonsignificant C $\times$ G was not an unexpected finding.

Within Policy Comparisons. The main effect of Traits and the two interactions, Traits by Conditions and Traits by Groups, were all significant, but no significant triple interaction (Traits by Conditions by Groups) was found. The significant Traits effect ($p < .001$) implies that the raters placed differential emphasis upon the six job traits as measured by the differences in the average validity coefficients. It would further imply that the raters were able to discriminate, and indeed did discriminate, between the job traits.

The significant Trait by Condition interaction ($p < .001$) is probably the most critical index, since it means that dissimilar policies emerged from the three conditions, that is, the job traits received differential relative emphasis within the three conditions. In other words, it demonstrates the differential significant influence which the traits, under the different conditions, had upon the raters as reflected in their policies based upon validity coefficients.

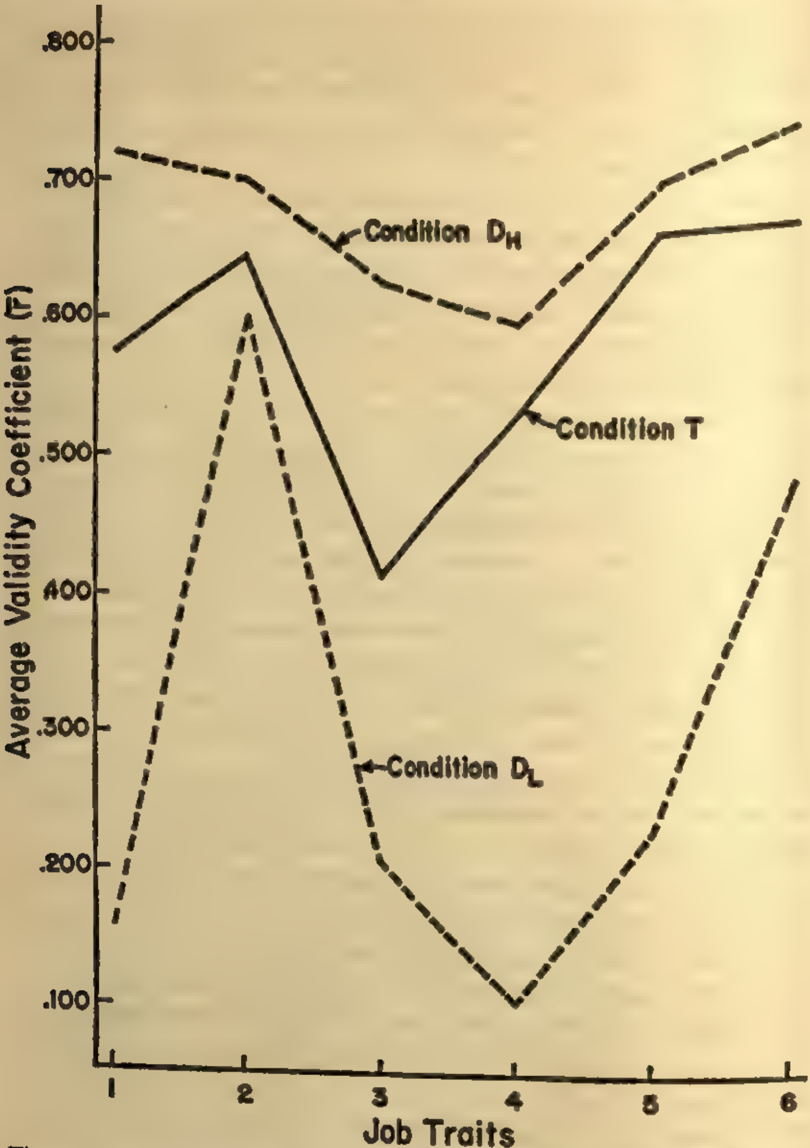


Figure 2. Average validity coefficients of the six predictor traits for each separate condition (single policy clustering stage).

This is shown most dramatically in Figure 2 where one can see that the profile patterns for the three different conditions differ greatly in their exaggeration even though the basic patterns are themselves similar. Thus the raters in Condition D_L, the orthogonal stimulus

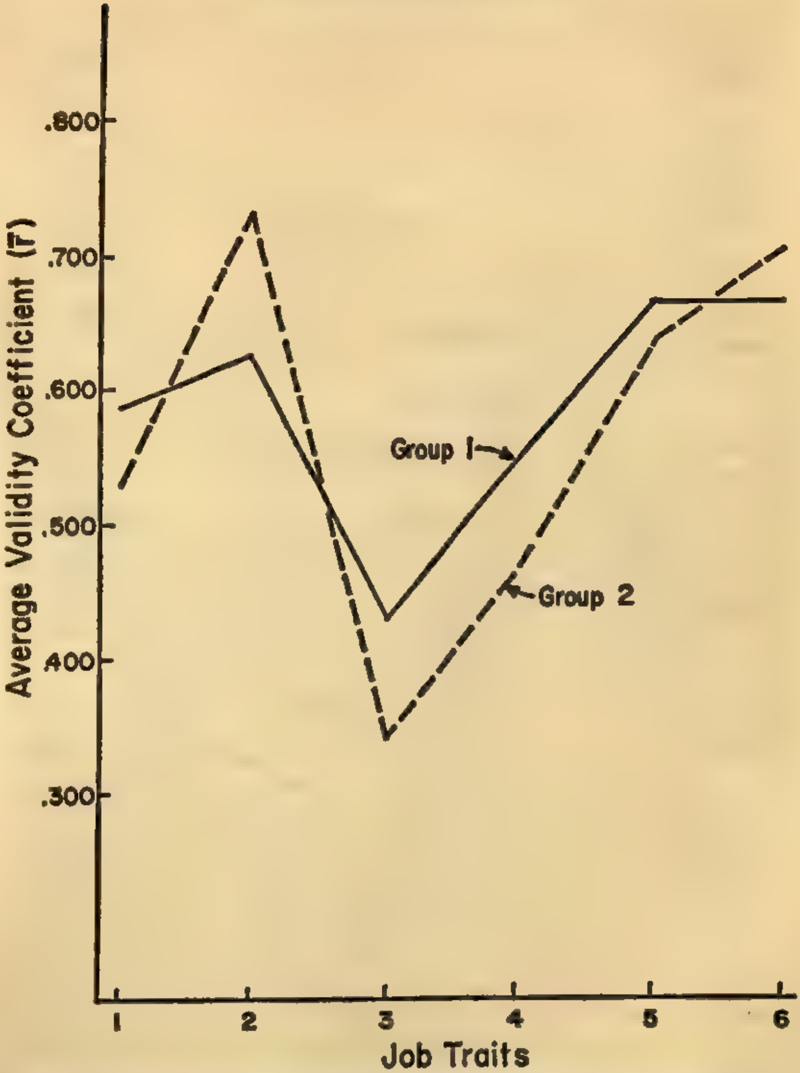


Figure 3. Average validity coefficients of the six predictor traits for Condition T raters (two subpolicy clustering stage).

dimension situation, exhibited much larger differences between the job traits than did the raters within Condition D_H, the highly correlated stimulus dimension situation, while it would appear that the Condition T raters fell somewhere between those of the previous two conditions.

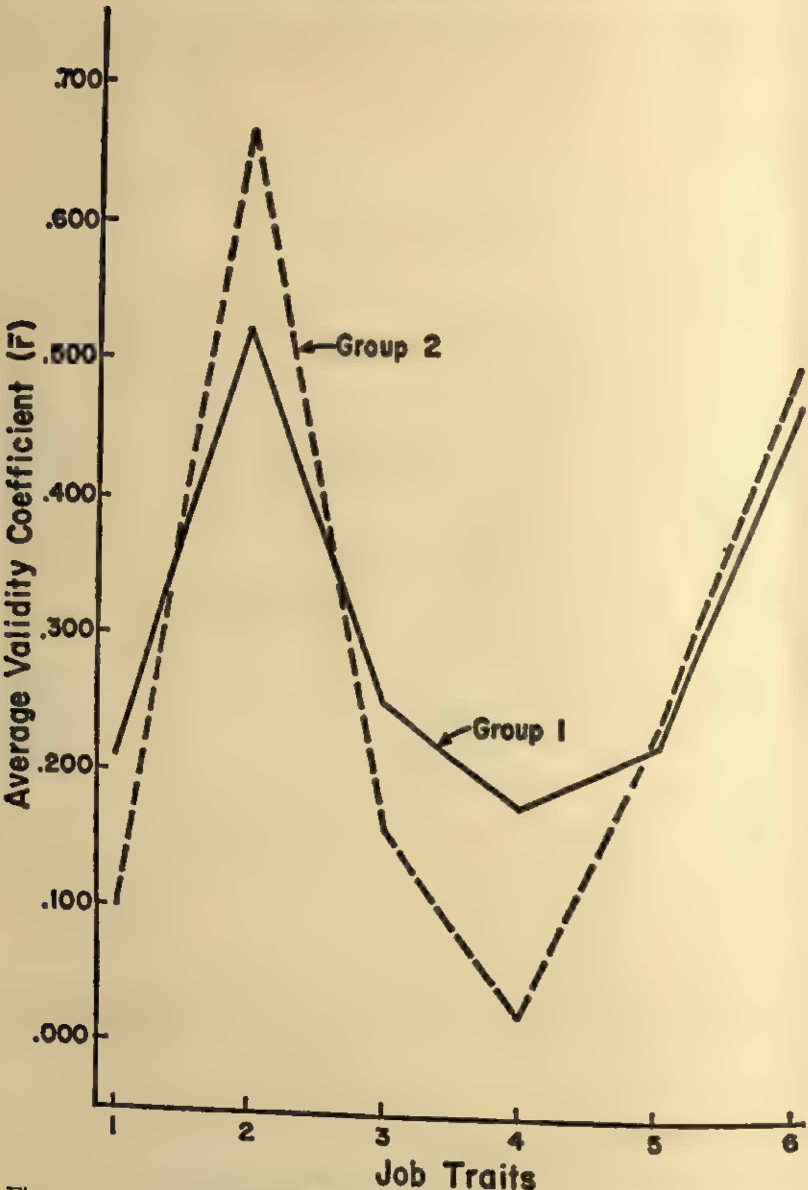


Figure 4. Average validity coefficients of the six predictor traits for Condition D_L raters (two subpolicy clustering stage).

The significant Trait by Group interaction ($p < .001$) indicates that the traits were differentially conceptualized within the two JAN subpolicy groups, but not also across the conditions as the

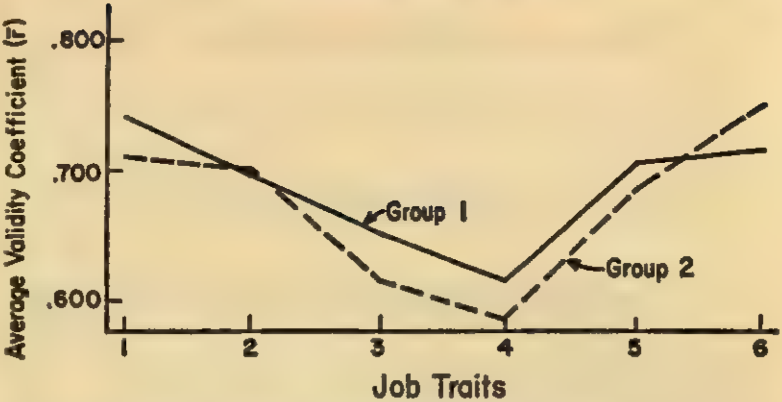


Figure 5. Average validity coefficients of the six predictor traits for Condition D_H raters (two subpolicy clustering stage).

case would have been had the triple (T $\times$ C $\times$ G) interaction been significant. From Figures 3–5 it can be seen that although disparate rater subpolicies exist for the two groups (significant T $\times$ G interaction), a similar pair of JAN subpolicy groups emerged in each of the three conditions; thus, the nonsignificant T $\times$ C $\times$ G interaction.

PROF Analysis

Finally, the PROF procedure (Wherry and Naylor, 1966) was employed on all profiles. This involved a principal axis factor analysis followed by a varimax rotation using the *R* matrix obtained by intercorrelating the validity vector for each rater with the validity vectors of all other raters across the combined three conditions.

There were five factors which emerged from the factor analysis. Examination of these factors indicated that the first four factors could probably best be interpreted as "condition" factors. This is clearly seen by looking at Table 9, which was developed by rank ordering the twenty highest factor loadings for each factor regardless of the condition. The results show that factors 1–4 are essentially "condition" factors in that they were defined almost exclusively by raters from conditions D_L, D_H, T, and again D_H respectively. Factor E was a hybrid. The logical implication being that the strategies obtained under the different conditions are not similar—rather they tend to represent independent policies.

TABLE 9

Twenty Highest Loadings for Each Factor

	Factor A			Factor B			Factor C			Factor D			Factor E		
	Load.	Rater	Cond.	Load.	Rater	Cond.	Load.	Rater	Cond.	Load.	Rater	Cond.	Load.	Rater	Cond.
1	994	14	D_L	990	24	D_H	962	30	T	938	9	D_H	950	36	D_L
2	991	37	D_L	947	35	D_H	958	11	T	889	26	D_H	810	16	T
3	988	32	D_L	946	34	T	956	17	T	847	31	D_H	561	13	D_L
4	982	16	D_L	945	36	D_H	888	28	T	828	33	D_H	524	24	T
5	976	22	D_L	945	2	T	878	5	T	795	38	D_H	497	12	D_H
6	973	7	D_L	940	6	D_H	823	27	T	780	1	D_H	394	33	T
7	971	33	D_L	937	29	D_H	813	7	T	764	12	D_H	389	25	D_H
8	968	31	D_L	936	18	T	751	12	T	630	20	D_H	383	38	D_H
9	963	15	D_H	933	2	D_H	749	21	T	530	25	D_H	375	13	T
10	962	10	D_L	924	19	D_H	734	19	T	502	27	T	371	14	T
11	956	21	D_L	905	22	D_H	703	13	T	450	7	D_H	360	30	D_L
12	946	27	D_L	882	32	T	700	22	T	421	8	D_H	348	7	T
13	946	23	D_L	854	21	D_H	673	20	T	399	30	D_H	343	19	D_L
14	945	34	D_L	851	16	D_H	672	8	T	384	18	D_H	316	1	D_L
15	942	28	D_L	851	16	D_H	669	14	T	362	17	D_L	298	18	D_L
16	937	9	D_L	830	12	D_L	641	25	T	353	17	D_H	292	29	D_H
17	934	4	D_L	827	24	T	622	33	T	307	22	D_H	284	22	T
18	922	15	T	825	7	D_H	611	10	T	300	13	D_H	281	17	T
19	914	5	D_L	818	28	D_H	605	36	T	295	10	D_H	270	4	D_L
20	906	3	D_H	812	31	T	563	37	T	293	10	T	269	31	T
	No. of Raters		Cond.	No. of Raters		Cond.	No. of Raters		Cond.	No. of Raters		Cond.	No. of Raters		Cond.
	17		D_L	13		D_H	20		T	17		D_H	9		T
	2		D_H	6		T				2		T	7		D_L
	1		T	1		D_L				1		D_L	4		D_H

Discussion and Conclusions

The results generally support the contention that variations in, or deviations from, a TRUE cue R matrix can and do significantly affect a rater's performance. This conclusion has important implications, particularly for those situations where artificial, rather than "real" stimulus objects are employed, or indeed, *must* be employed.

The most impressive evidence which supports the conclusion that deviations from a TRUE cue R matrix do significantly affect a rater's judgment policy evolved from the analysis of variance, in that the Trait by Condition interaction was found to be significant. This significant T $\times$ C interaction implies that the rater policies, measured by validity coefficients, are different in different ways within the three conditions, that is, that the policy profiles across the six traits for each condition departed significantly from parallelism. This conclusion is further supported by the interpretations attached to the significant main effects of Conditions, which implies that the policy *levels* for the three conditions, measured by the magnitudes of the average validities for the traits, are different. Thus both profile pattern and profile level seem to be affected by R matrix structure.

Further evidence of the effect of Conditions upon judges' performance can be seen from the behavior of other indices, such as R^2 and R_c^2. For instance, it is evident that the stimuli used in all conditions resulted in highly consistent intrajudge performance as depicted by the large R^2 values for all raters, which in turn, denote a high degree of success in capturing rater's policies. However, despite the fact that most R^2 values were large, there exist systematic differences between these values across the three conditions. These differences may have resulted from the fact that while raters within the orthogonal condition (D_L) were able to discriminate clearly and meaningfully between the predictor variables (job traits), at the same time they found the decision making task difficult since each trait must be attended to separately. This would cause the rater R^2 values to drop, as would be expected. On the other hand, the raters within the highly correlated condition (D_H) were less able to discriminate, but found the decision making task easier, resulting in high rater R^2 values.

Also, the composite R^2 values (R_c^2) obtained from the sequential stages of JAN were found to be relatively large indicating that even when all the raters within a condition were clustered into a single composite, a high level of predictability remained. In addition, it implies the individual raters' policies within a condition were relatively homogeneous, a fact borne out by the factor analysis. However, the systematic differences between R_c^2 values across conditions simply reflect the difference previously explained pertaining to the individual raters' R^2 values.

The PROF factor analysis also yielded rather conclusive data supporting the effect of R matrix distortion. The factor analysis as performed had the chief advantage that the similarities or dissimilarities of factor-strategies across conditions could emerge, since all raters were included in a single analysis. This allowed for the likelihood of obtaining "condition" factors, as was borne out in the results. Indeed, the first four factors were clearly defined by raters from specific conditions.

The implications of these data seem rather straightforward. If one is interested in obtaining information about the judgmental policies of individuals toward a certain class of stimulus objects he needs to be certain that the underlying cue R matrix for his experimental stimuli is representative of the R matrix describing the population of stimuli of which his sample is assumed to be a subset. Otherwise, his ability to generalize is obviously going to be limited.

REFERENCES

Ash, P. The SRA Employee Inventory—A Statistical Analysis. *Personnel Psychology*, 1954, 7, 337–364.

Baehr, M. E. A Factorial Study of the SRA Employee Inventory. *Personnel Psychology*, 1954, 7, 319–336.

Bottenberg, R. A. and Christal, R. E. *An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency*. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Division, March 1961 (WADD-TN-61-30, ASTIA Document AD-261 615).

Naylor, J. C. and Wherry, R. J., Sr. *Feasibility of Distinguishing Supervisor's Policies in Evaluation of Subordinates by Using Ratings of Simulated Job Incumbents*. Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, October, 1964. (PRL-TR-64-25).

Naylor, J. C. and Wherry, R. J., Sr. The Use of Simulated Stimuli,

Multiple Regression, and the JAN Technique to Capture and Cluster the Policies of Raters. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1965, 25, 969–986.

Wherry, R. J., Sr. An Orthogonal Re-rotation of the Baehr and Ash Studies of the SRA Employee Inventory. *Personnel Psychology*, 1954, 7, 365–380.

Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., and Fallis, R. F. Generating Multiple Samples of Multivariate Data with Arbitrary Population Parameters. *Psychometrika,* 1965, 30, 303–313.

Wherry, R. J., Sr. and Naylor, J. C. Comparisons of Two Approaches—JAN and RPOF—for Capturing Rater Strategies. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 267–286.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 605-616.

THE TEST-RETEST RELIABILITY OF CHILDREN'S RATINGS ON THE SEMANTIC DIFFERENTIAL[1]

FRANCIS J. DI VESTA

AND

WALTER DICK[2]

The Pennsylvania State University

THE present study of the reliability of the semantic differential technique was made to supplement the data from recent investigations (Di Vesta, 1964) providing positive evidence regarding its use with children. Several kinds of reliability might be adequately studied. Among these are item, factor-score, and concept-meaning stability. Osgood and his associates (1957) report highly acceptable levels of these various forms of reliability when the technique is used with adults. Jenkins, Russell, and Suci (1958) also report coefficients of stability of .90 or higher for mean scale values and mean profiles of concepts based on ratings made by 20 or more college students. However, comparable studies of semantic differential reliability, when used with children, are not known to the authors. Data regarding the reliability at several grade levels should provide useful guides to other investigators in determining the adequacy or limitations of the semantic differential for applications with individuals or groups of children in the elementary grades.

Since the semantic differential yields a profile of ratings, a number of different scores may be used in determining the reliability of

---

[1] The study reported here was supported by Grant Number MH2900 from the National Institute of Mental Health and Grant Number HD-00872 from the National Institute of Child Health and Human Development, United States Public Health Service. The data were collected at Syracuse University. All computations were made by the Pennsylvania State University Computing Center.

[2] Now at Florida State University.

this instrument. Its stability can be determined by the test-retest reliability coefficients calculated for scale items, factor scores, distances (D_c) between concepts, and polarity, that is, distances (D_p) of ratings from the origin of the semantic space. These, in turn, may be based on individual or mean ratings. Since the reliability of distances between concepts is dependent on the reliability of the factor scores on which they are made, the decision was made to limit the present analysis to test-retest reliability of scales, to the deviation of factor scores from pretest to posttest, and, where appropriate, to test-retest reliability of factor and polarity scores, separately based on individuals and means. The reliability of the semantic differential was studied separately for delayed and immediate retests.

Method

The data for the present study were gathered as part of three factor analytic studies of the development of children's meaning. Retest data were gathered in two of these three independent studies. Since the subjects, selection of scales, selection of concepts, and method have been described in detail in an earlier publication (Di Vesta, 1964) only the procedure related to the reliability study will be presented here.

In both studies, the reliability of individual scales and of factor scores are presented. Although a number of scales were aligned on each factor, only those two scales with the highest loadings were averaged to obtain each factor score. The factors, and the scales by which they were defined, are as follows: Evaluation . . . *good-bad* and *friendly-unfriendly;* Potency . . . *strong-weak* and *brave-not brave;* Activity . . . *fast-slow* and *moving-still;* Size . . . *big-small* and *long-short;* and Warmth . . . *hot-cold* and *wet-dry*. These factors were identified through rotation of the principal components factors by the Equamax routine to achieve simple structure.

Delayed Retest

In the main study (identified as Study II, in the earlier report, of which the delayed retest study was a part), 100 *S*s in each of the grades two through seven rated 100 concepts on 27 scales. Each *S* rated 20 concepts. Except for ratings made by *S*s in the seventh

grade, all ratings were over a three- or four-week period in order to accommodate the convenience of the school personnel. The ratings by the seventh grade *S*s were made during a three-day period. After a one-month interval, a graduate assistant asked each *S* in each class to rate again *one* of the concepts (selected at random) previously rated. Each *S* rated a concept different from that of the other *S*s within his group.

A total of 522 *S*s are represented in this study: in grade two, $N = 86$; in grade three, $N = 87$; in grade four, $N = 91$; in grade five, $N = 73$; in grade six, $N = 89$; and in grade seven, $N = 96$. The number of concepts rated corresponds to the number of *S*s in each grade. Although more than 100 *S*s were originally tested, something less than this number was finally obtained because of absences during either the test or retest period. In addition, a few papers were rejected in each class because of incomplete responses or because the papers were obviously pattern-marked.

Immediate Retest

In the main study in which the immediate retest was administered (identified as Study III in the earlier report) 100 additional concepts in sets of 20 concepts were rated by *S*s in grade three ($N = 194$); grade five ($N = 129$); and grade seven ($N = 181$). These *S*s were in a different school system than those used for the delayed retest. Each *S* rated 20 concepts, selected at random, on 21 scales, in two series of ten concepts each, during a period of three to five days. The eleventh concept in the second series was a repeat of one, selected at random, previously rated in *that* series. The concept in the retest was administered separately from the other concepts. Some *S*s recognized that the concept had been previously rated, but were always instructed that the rating was to be made independently of the earlier rating. Each of the 50 concepts was rated by three to five *S*s. Thus, although the *N* was small, we were able to use mean scores as well as individual scores for the analysis. As a result of absences at the time the original test or the retest was administered, a total of 488 *S*s were actually used in the immediate retest study: $N = 192$ in grade three, $N = 123$ in grade five, and $N = 173$ in grade seven. Scores were obtained on 50 concepts but because of absences one concept is not represented in each of the grades three and five.

Results

Delayed Retest

The data for this study are based entirely on comparisons of individuals. Each individual's rating is of a concept different from all other individuals in his group. The interval of time varied from the original (test) to second (retest) rating of the concept. In general the period between the two ratings was roughly four weeks.

The individual test-retest scores on each scale were correlated. In Table 1 it may be seen that all of the correlation for scales are

TABLE 1

Correlations between Test-Retest Scale Ratings Made by Individuals: Delayed Retest

Scales	Grades					
	2	3	4	5	6	7
Wrong–Right	.34	.37	.68	.66	.59	.72
Round–Square	.31	.33	.34	.34	.46	.54
Sweet–Sour	.22	.41	.52	.57	.77	.71
Dark–Light	.49	−.09	.12	.60	.42	.51
Little–Big	.53	.37	.39	.46	.39	.66
First–Last	.18	.22	.20	.43	.37	.65
Wet–Dry	.33	.12	.42	.71	.39	.45
Not Brave–Brave	.39	.45	.41	.25	.38	.54
Ugly–Pretty	.13	.55	.47	.65	.52	.56
Light–Heavy	.47	.34	.40	.39	.47	.63
Make Believe–Real	.33	.34	.48	.25	.35	.59
Sad–Funny	.19	.39	.54	.65	.45	.63
Still–Moving	.31	.35	.57	.66	.41	.48
Good–Bad	.33	.39	.69	.85	.68	.75
Quiet–Loud	.50	.26	.56	.38	.46	.68
Cold–Hot	.34	.54	.10	.21	.39	.31
Smooth–Rough	.25	.29	.45	.44	.52	.58
Friendly–Unfriendly	.41	.33	.62	.69	.68	.63
Strong–Weak	.32	.25	.54	.53	.73	.49
Red–Blue	.46	.13	.21	.51	.46	.47
Tight–Loose	.25	.13	.42	.61	.19	.14
New–Old	.47	.28	.31	.64	.31	.65
Fast–Slow	.34	.23	.35	.31	.56	.51
Long–Short	.15	.36	.30	.33	.31	.47
Soft–Hard	.34	.29	.62	.66	.45	.54
Dull–Sharp	.30	.47	.29	.12	.27	.45
Same–Different	.20	.09	.43	.31	.16	.46

low, ranging from an average correlation, across grades, of .27 to .56. A progressive increase in the magnitudes of the coefficients for grades averaged over scales is reflected in the range, from an

average of .33 in the second grade to an average of .55 for the seventh grade. The largest correlation coefficients are to be found for the Evaluation scales followed in magnitude by the scales for Potency and Activity. The lowest correlations are for those scales in the Warmth factor and a composite of other scales including *same-different*, *dull-sharp*, and *tight-loose*.

A second analysis was based on the factor scores obtained by simply adding the ratings on two scales associated with each factor. Two dominant trends are apparent in these data as presented in Table 2. The first is a definite increase in reliability of ratings made on the semantic differential between the third and fourth grades. The second trend is that the ratings made of the Evaluation, Potency and Activity factors are, in general, more reliable than any of the remaining factors.

TABLE 2

Correlations between Test-Retest Factor Scores of Ratings Made by Individuals: Delayed Retest

Factors	Grades 2	3	4	5	6	7
Evaluation	.39	.49	.75	.86	.79	.78
Potency	.55	.49	.60	.40	.64	.55
Activity	.51	.35	.54	.64	.64	.54
Size	.19	.43	.54	.54	.46	.63
Warmth	.41	.26	.21	.61	.55	.22

A measure of meaning related to the factor scores is the intensity of meaning. However, rather than employing a single score, the intensity of meaning (polarity) is determined by using the deviation of the profile or factor scores from the origin of semantic space and is calculated by the formula $D_{il} = \sqrt{\Sigma_j d_{il}^2}$ where i is the factor score, j is the factor and l is, in this case, the constant 4 (the mid-point, or origin, of the scale). The scores for each of the five factors were used in obtaining the polarity index (D_p). The reliabilities of the D_p scores are presented in Table 3. Since the polarity measure, as an index of meaning, is of theoretical significance the means and standard deviations of the Ds are also presented in this table. As in the previous analysis the higher reliability coefficients are found in the upper three grades. Overall, the stability of the polarity measure is somewhat higher than for factor or scale scores, and gradually increases from the second to the seventh grade. The mean polarity

TABLE 3

Correlations, Means, and Standard Deviations of Test-Retest Polarity Scores of Ratings Made by Individuals: Delayed Retest

Grade	Coefficient of Reliability	Pre-Test		Post-Test	
		Mean	SD	Mean	SD
2	.50	3.48	1.18	3.17	1.47
3	.55	3.11	1.10	3.03	1.10
4	.45	3.10	1.06	2.71	1.32
5	.62	3.20	1.23	3.03	1.11
6	.62	3.29	1.10	3.14	1.16
7	.73	3.24	1.20	3.01	1.11

index is lower on the retest at every grade-level than it is on the original test.

Another indication of reliability of interest in the application of the semantic differential is the determination of the probability of obtaining given absolute deviations of factor scores from test to retest. The percentages of responses giving each deviation (*per cent*) and associated probability (*P*) were computed for five factor-score averages by grade level.[2] Since each score was obtained by averaging over two scales, deviations were in terms of half scale units. There were few appreciable differences among grades in these data. In general, the results correspond to those presented by Osgood, Suci and Tannenbaum (1957) who used a larger number of scales in obtaining factor scores.

Immediate Retest

The relatively low reliability coefficients for the first study may be attributed to the interval of four weeks between test and retest and to the fact that only individual ratings were used. These considerations are of importance since children's concepts are continually undergoing change and instruction in school may have an appreciable influence on the ratings. In addition, data from correlated scales of the semantic differential are generally pooled to obtain factor scores which represent the judgments of individuals or groups. The present analysis was undertaken to determine the effects of combining scale scores compared to the effect of using in-

[2] The deviations and associated probabilities for the delayed retest and immediate retest studies are presented in Tables A and B. They have been deposited with the American Documentation Institute. Order Document No. 8900, remitting $1.25 for 35-mm. microfilm or $1.25 for 6 by 8 in. photocopies.

dividual scores as in the previous analysis. This study also attempts to examine the stability of the semantic differential over a short period of time.

While our procedures permitted us to obtain concept means based on groups of three to five individuals, it still is not the most optimal procedure. In studies (Jenkins, Russell and Suci, 1958; Miron, 1961) employing adult *S*s, such group means are based on 20 or more *S*s. Hence, the estimates of reliability presented here must be considered as conservative estimats of the stability of children's ratings. The period between the test and retest was one or two days. All methods of analyzing data for the first (delayed retest) study were followed in the present study. Judgments of concepts were made on 21 scales (shown in Table 5) rather than 27 as in the first study. The five factors of Evaluation, Potency, Activity, Size, and Warmth are defined by the same scales described above and so are directly comparable.

The correlations of test-retest ratings on each scale over all individuals and concepts within a grade are presented in Table 4.

TABLE 4
Correlations between Test-Retest Scale Ratings Made by Individuals: Immediate Retest

	Grades		
Scales	3	5	7
Heavy–Light	.74	.72	.86
Quiet–Noisy	.66	.64	.79
Dry–Wet	.42	.60	.67
Hot–Cold	.38	.35	.63
Make Believe–Real	.52	.65	.52
Terrible–Wonderful	.65	.74	.76
Smooth–Rough	.62	.49	.62
Powerless–Powerful	.55	.70	.65
Usual–Strange	.50	.51	.72
Unfriendly–Friendly	.62	.53	.69
Moving–Still	.56	.60	.66
Strong–Weak	.63	.53	.60
Good–Bad	.74	.64	.81
Fast–Slow	.43	.39	.58
Old–New	.32	.50	.66
Soft–Hard	.62	.66	.79
Little–Big	.58	.69	.71
Straight–Curved	.70	.58	.61
Dark–Light	.41	.35	.62
Long–Short	.39	.34	.53
Brave–Not Brave	.67	.70	.61

Compared to the delayed retest data, reported above, all coefficients are considerably higher in the present study. The coefficients for grades, averaged over scales, are .56 for each of the grades three and five, and .67 for grade seven; the coefficients for scales (averaged across grades) range from .42 to .77. Relative to the magnitude of the coefficients for the other factors those for Evaluation, Potency, and Activity are found to be the most reliable ratings; and judgments made on the *hot-cold, old-new, dark-light* and *long-short* scales are less reliable. These results are consistent with those reported for the delayed retest.

The factor scores were averaged over ratings, made by groups of three to five *S*s, of each concept thus yielding factor scores for concepts. Correlations between concept scores on the test and retest administrations were then obtained. The factor scores for each individual were also correlated, as in the delayed retest study. These data are presented in Table 5. All correlations for concepts are of

TABLE 5

Correlations between Test-Retest Factor Scores of Individuals and Concepts (Groups): Immediate Retest

Factor	Grade 3		Grade 5		Grade 7	
	Concepts	Individuals	Concepts	Individuals	Concepts	Individuals
Evaluation	.89	.77	.87	.78	.94	.84
Potency	.79	.67	.79	.69	.84	.72
Activity	.80	.62	.81	.64	.73	.69
Size	.82	.64	.67	.64	.89	.75
Warmth	.60	.38	.70	.64	.85	.68

acceptable magnitude and in no instance was the correlation for concepts lower than the respective correlation for individuals. All correlations are sufficiently high to indicate the functional utility of separate factor scores, based on only two scales, for research purposes.

The reliabilities of polarity scores were also obtained for both individuals and concepts. The coefficients of reliability, thus determined are shown in Table 6. In that display, it may be seen that the polarity of concepts for individuals is relatively more stable during short intervals than found in the first study. Not surprisingly, the correlations for concepts are higher than for individuals.

TABLE 6

Correlations, Means, and Standard Deviations of Test-Retest Polarity Scores of Ratings Based on Individuals and Concepts (Groups): Immediate Retest

	Coefficient of Reliability	Pre-Test		Post-Test	
		Mean	SD	Mean	SD
Concepts					
Grade 3	.82	3.70	1.14	3.35	0.94
Grade 5	.69	3.91	1.19	3.72	1.21
Grade 7	.89	3.85	1.12	3.83	1.11
Individuals					
Grade 3	.74	4.65	1.52	4.53	1.53
Grade 5	.67	4.58	1.21	4.54	1.29
Grade 7	.79	4.53	1.22	4.54	1.27

The tendency, noted earlier, for the polarity of judgments to decrease on the second testing prevails. However, a small general increase in polarity (meaning) with age in the concept means is also to be noted.

Absolute deviations from the test and retest administrations were also computed. The percentages of individuals whose ratings did not change from test to retest were substantially higher in the "no deviation" range than in the delayed retest study. Otherwise, there were no appreciable differences among grades. However, even with these data we can expect about 10 to 15 per cent of the ratings to deviate by more than two scale units.

Discussion

The findings in the present study clearly suggest that over brief periods, the semantic differential is a reliable instrument when used with children. The coefficients of stability are comparable to those ordinarily obtained by other techniques requiring judgments by individuals. In addition, it can be used with children as young as those in the third grade, and perhaps those in the second grade with little change in the procedures used in administering the scales to adults. However, with children in the lower grades it is advisable to use larger numbers of *S*s and, perhaps, three or more scales in arriving at factor scores. The reliability of the data generated by the semantic differential is higher when ratings from two scales are combined to obtain factor scores than when only single scales are used. There is considerable variability in the reliabilities of individ-

ual scales and in many cases they are unsatisfactory. The extent to which such low coefficients reflect instability of the measure or of the concepts being judged can be determined only by further investigations. Warr and Knapper (1965), for example, found that test-retest reliability depends on the concept being judged, especially where metaphorical extensions are involved. For usual research practice, where the concern is with only one to three dimensions, it seems likely that the investigator will choose to combine the scores from three to five scales in order to realize higher stability.

As with even the most standardized psychometric instruments, the stability of the ratings on the semantic differential is greatly reduced when the interval between test and retest is increased. This seems especially so for children in the second and third grade. Whether instability in the lower grades, over a one-month interval, is due to real change in concept meaning or merely reflects error variance must await further investigation, the procedures for which are suggested by Osgood *et al.* (1957). However, the results of our second study suggest that the instability of ratings by children in the lower grades may be the result of actual change in concept meaning, that is, the ratings by third grade children are very stable when test and retest administrations are spaced one to two days apart.

The reliability of the ratings of individual scales, and of the factor scores, also indicates the saliency of the familiar Evaluation, Potency and Activity dimensions. These appear to be general and pervasive factors. It seems likely that they are experienced early in the development of the child. That other, somewhat less salient and less pervasive, factors exist in the child's language has been demonstrated in our earlier studies. That they can be measured reliably is demonstrated in the present study.

The absolute deviations of the factor scores from test to retest are approximately the same as those found with adults by Osgood, *et al.* (1957) with little appreciable difference among grades. However, the probability of a change in factor scores of more than 2.00 scale units would occur 10 to 15 per cent of the time with the individual ratings of children compared to a probability of .05 for a change in factor scores of more than 1.00 to 1.50 for data based on individual adult ratings where the test-retest interval was thirty minutes.

Summary and Conclusions

The stability of the semantic differential technique was examined under delayed and immediate test-retest conditions. The study was conducted with children in grades two through seven in the former and in grades three, five, and seven in the latter condition. There were 522 *S*s in the delayed retest study and 488 *S*s in the immediate retest study.

The coefficients of stability were based upon the correlation of factor scores derived by simply averaging the ratings of two scales with high loadings on each of five factors in the delayed retest study and seven factors in the immediate retest study. In general the semantic differential technique was shown to be an acceptably stable instrument when used with children as young as the third grade under immediate retest conditions. Reliability for concepts based on group means is higher than that for individuals even when groups are composed of only three to five subjects. An extrapolation of these results suggests the possibility that levels of reliability of concept scores of children will approach those obtained for adults if 15 to 20 *S*s are used in the rating procedures. The lower reliability coefficients obtained for the delayed retest condition were attributed in part to changes in meaning of concepts although the existence of larger measurement errors was also considered as a possible contributing factor to the lower reliability of ratings made by second and third grade children.

The individual scales and the factor scores relating to the salient dimensions of Evaluation, Potency, and Activity were found to be more reliable than the remaining scales under both delayed and immediate retest conditions. Used with caution, the factor scores of Size, and Warmth were sufficiently reliable for research purposes and suggest other dimensions for further investigations in which the semantic differential technique is employed.

REFERENCES

Di Vesta, F. J. A Developmental Study of the Semantic Structures of Children. Technical Report No. 8. Research Grant No. HD-00872, National Institute of Child Health and Development, 1964.

Jenkins, J. J., Russell, W. A., and Suci, G. J. An Atlas of Semantic

Profiles for 360 Words. *American Journal of Psychology,* 1958, 71, 688–699.

Miron, M. S. The Influence of Instructional Modification upon Test-retest Reliabilities of the Semantic Differential. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 883–893.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. *The Measurement of Meaning*. Urbana, Illinois: University of Illinois Press, 1957.

Warr, P. B. and Knapper, C. Some Factors Affecting the Reliability and Validity of Semantic Differential Scales. Aberdeen, England: Paper read to the British Psychological Society, April, 1965.

Educational and Psychological Measurement
1966, 26, 617-622.

EVIDENCE ON PROBLEMS IN ESTIMATING COMMON FACTOR SCORES

JOHN L. HORN and WILBUR C. MILLER
University of Denver

With the advent of the computer in psychological research, problems in estimating factor scores have come increasingly to the fore. In a recent article Horn (1965) compared various such estimation methods in terms of their theoretical-mathematical properties and by intercorrelating the scores estimated by the different procedures. His findings showed that the scores estimated by three complete methods were highly similar, as were those estimated by three so-called incomplete methods, but that these two broad classes of techniques yielded scores which correlated only .80 on the average. He found, also, that the intercorrelations among different orthogonal factors estimated by each of the three complete methods averaged .08 (zero being the desired value), whereas these intercorrelations for the incomplete methods averaged .14.

This kind of information no doubt has relevance for questions concerning the usefulness of these various techniques in research and applied work. But an essential item of information is lacking. It is important to know not only how the factors estimated by the different procedures are interrelated among themselves, but also how each relates to the variables upon which a factor analysis is based. One purpose of the present investigation is to supply this kind of information.

The data upon which Horn's (1965) analyses were based were somewhat atypical. At least they were not representative of those found in studies where a strong positive manifold prevails among the correlations for variables. On first consideration this might seem to be a rather minor matter. But when one stops to take ac-

count of the fact that (at least with some methods of estimation) this can imply that all variables will be given positive weights in the equations for estimating two supposedly independent factors, then it becomes apparent that with certain kinds of data these methods could yield factor estimates which were much more highly intercorrelated than would be implied by the factor solution. For example, if factor scores are estimated by adding standard scores (for all variables) weighted by the factor loadings of the variables (Horn's method 4) and the factor loadings are all positive for two factors (these being cooperative: cf. Cattell (1957)), then all that distinguishes the two sets of factor scores is the *pattern* of positive differential weights. But if there are many variables, then, as Richardson (1941) showed analytically and as Guilford, Lovell and William (1942) found empirically, these two linear composites are apt to correlate quite highly (i.e., unless the patterns of weights are very different indeed). Cooperative factors are not likely to be defined by widely different coefficients. Hence under these conditions factor score estimates are not likely to be as independent as the factor solution implies.

Independence would tend to be maintained with the complete methods, on the other hand, since in these cases, as in the usual multiple regression equation, if one variable of a highly correlated pair is given a positive weight, the other is likely to be given a widely different negative weight.

Horn's (1965) data did not show the high degree of positive manifold here stipulated. It is not surprising, therefore, that he did not find this effect. The intercorrelations among different factors were low for all of the factor estimation methods he studied. But would these correlations be low in all cases if a high degree of positive manifold obtained among the intercorrelations for variables? A second purpose of the present investigation is to examine this question by analyses with data having the characteristics here specified.

Procedures and Results

A series of 20 teacher ratings was obtained on a sample of 195 children. Correlations between these scales were determined on a subsample of 167 children for which complete data were available. These scales, although measuring what were supposed to be inde-

pendent attributes—such as physical skill, creativity, leadership ability, sense of responsibility, etc.—nevertheless yielded correlations showing a very substantial positive manifold (this, of course, reflecting in part the halo phenomenon so often observed in behavior ratings).[1] Five principal axes factors were found to account for the major portion of the reliable, common variance among these correlations. These factors were rotated to a position satisfying Kaiser's (1958) Varimax criteria. Then, using the total sample of 195 children, the following procedures for estimating factor scores were employed:

$$\begin{aligned}
&(1) \quad F_1 = ZR^{-1}A \\
&(2) \quad F_7 = ZR^{-1}B \\
&(3) \quad F_4 = ZA \\
&(4) \quad F_6 = ZB
\end{aligned}$$

where the F_j (subscripts chosen to be consistent with Horn, 1965) are N by m matrices of factor scores, Z is the N by n matrix of original standard scores, R^{-1} is the n by n inverse of the matrix of intercorrelations among variables, A is an n by m matrix of factor coefficients (i.e., a structure) and B is an n by m matrix in which unity is substituted for each "salient" coefficient in A and zero is substituted for every other loading. (A variable was defined as salient for a factor if it had its highest loading on that factor. In all cases this meant that the salient had a loading of at least .45 on the factor in question and frequently the loadings were higher than this. A variable was defined as salient [in this sense] for one and only one factor.) The factor scores estimated in these ways were intercorrelated, whence the results summarized in Table 1 were abstracted.

Table 1(a) shows results comparable to those obtained by Horn (1965). That is, Horn had found—with his larger samples of variables—that the average correlation for the same factor estimated by procedures F_1 and F_6 was .79; for procedures F_1 and F_4 this was .82 and for procedures F_4 and F_6 it was .90 (procedure F_7 was not

[1] These intercorrelations, the factor structure (pattern) matrix and sets of intercorrelations to be mentioned in later sections of this paper, have been deposited with the American Documentation Institute. They can be obtained by remitting $1.25 for microfilm or $1.25 for photoprints (check payable to Chief, Photoduplication Service, Library of Congress) for Document No. 8909, ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540.

TABLE 1

Average Correlations between Factor Scores Estimated by Various Procedures

(a)
For the Same Factors Estimated by Different Procedures

	Procedure F_7	F_4	F_6
F_1	.656	.651	.818
F_7		.259	.572
F_4			.918

(b)
For Different Factors Estimated by the Same Procedure

Procedure F_1	F_7	F_4	F_6
.127	−.217	.902	.642

used); in the present study these values are .82, .65 and .92 respectively. But in Table 1(b) results quite in contrast to the earlier findings are shown. The correlation between different, supposedly independent, factors estimated by the F_4 and F_6 procedures are much higher than would be desirable in most studies. This outcome results because all the variables are positively correlated and the weights which are assigned in the estimation equations are not widely different. As noted before, this outcome is predicted by Richardson's (1941) early theoretical analysis. It agrees with the results obtained by Guilford, *et al.* (1942).

TABLE 2

Average Discrepancies between Correlations (For Variables and Factor Score Estimates) and Factor Structure Coefficients (Loadings)

	Factor Score Estimation Procedure			
	F_1	F_7	F_4	F_6
Average for:				
1. Salient Variables	.104	−.123	.147	.204
2. Non-salient Variables	.089	−.218	.434	.297

Table 2 gives a summary of results which permit of some comparisons between estimation procedures in terms of the accuracy with which the method gives scores that correlate with variables in the way suggested by a factorial solution. A discrepancy larger than zero indicates a failure of the estimation procedure. Table 2 shows that F_1, F_4 and F_6 give factor scores that are somewhat

biased in the direction of having a higher correlation with variables than would be suggested by the factor structure, whereas the scores obtained by method F_7 correlate consistently lower than would be suggested. The results follow the same pattern for both salient and non-salient variables. Correlations which are larger than expected are no doubt so in part because of autocorrelation, since the variance of the variable is included in the factor score with which it is correlated. The smaller correlations resulting from use of method F_7 reflect the fact that in making the least squares estimation, suppression weights are built in (by taking the inverse of R). When these are used with all variables (and all factor structure coefficients), the resulting factors correlate as required with variables (if corrections for autocorrelation are made, the discrepancies for method F_1 reduce practically to zero). But when used with only a few of the variables in method F_7, over-suppression results and the factor score correlations with variables are thus reduced.

The general findings from this investigation thus argue for a position in opposition to Horn's (1965) implicit contention that there is relatively little to choose between the different procedures. If there is a rather solid positive manifold among the intercorrelations for a set of variables or if, under any circumstances, two factors are highly cooperative (even though independent), then use of method F_4 is likely to produce factor scores which are too highly intercorrelated. In these circumstances one of the complete procedures, as here represented by method F_1, should be used.

Summary

Four methods of estimating common-factor scores were compared for use in situations where consistent positive correlation obtains between variables. Analytic considerations suggested that under these conditions scores estimated by incomplete methods might show higher intercorrelations than would be desirable. This prediction was borne out by empirical analyses. In addition data were presented which allowed for comparison between factor estimation methods in terms of the correlations which the resulting factor scores had with variables. It was found that when the most common least-squares ("complete") procedure for estimating factor scores was used, the correlations between factor scores and variables

were very close to those shown in the factor structure matrix, but that when other ("incomplete") procedures were employed, the differences between the estimate-variable correlations and the factor structure coefficients were larger. These results thus argue for use of one of the complete procedures when this is possible.

REFERENCES

Cattell, R. B. *Personality and Motivation Structure and Measurement.* New York: World Book Company, 1957.

Guilford, J. P., Lovell, C. and William, R. M. Completely Weighted versus Unweighted Scoring in an Achievement Examination. Educational and Psychological Measurement, 1942, 2, 5–21.

Horn, J. L. An Empirical Comparison of Various Methods for Estimating Common Factor Scores. Educational and Psychological Measurement, 1965, 25, 313–322.

Kaiser, H. F. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika,* 1958, 23, 187–200.

Richardson, M. W. The Combination of Measures. In Horst, P. (Ed.). *The Prediction of Personal Adjustment.* New York: Social Science Research Council, 1941.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 623-631.

NOTE ON RANK BISERIAL CORRELATION[1]

GENE V GLASS
University of Illinois

ONE may obtain data to be correlated in many different forms. They may be in the form of observations on normally distributed variables, n consecutive ranks, a dichotomy, a trichotomy, etc. In this paper we shall be concerned with developing a measure of the correlation between one variable comprising n consecutive untied ranks and a second variable comprising a dichotomy. We shall regard the dichotomy (scored 0, 1) as having a ranking variable underlying it. In this regard it is analogous to the assumption made in biserial correlation that a normally distributed variable has been forced into a dichotomy. Of major interest is a surprising equivalence which results between the coefficient found in this paper and a coefficient due to Cureton (1956). It is to present and prove the equivalence of these two coefficients, which though quite different in origin and derivation are algebraically equivalent, that this paper is being written.

Derivation of the Rank-Biserial Correlation Coefficient, rb

We shall assume that Y is a variable composed of the n consecutive untied ranks, $1, 2, \ldots, n$. Let X' be a variable of the same type. A measure descriptive of the correlation between Y and X' is Spearman's rho, the product-moment correlation coefficient between the two variables.

Assume that we do not know X' but instead we have a (0, 1) dichotomous variable which has a ranking variable, X', underlying it. The n_0 persons at 0 on X are assumed to have the ranks 1 through

[1] I wish to thank Dr. David E. Wiley for his invaluable assistance at the early stages of the work reported in this paper.

n_0 on X'; the n_1 persons at 1 on X are assumed to have the ranks $n_0 + 1$ through n on X'.

One might wish to gain information about the population Spearman's rho between Y and X' from the scores on Y and X. There is an analogy between this situation and that in which the conventional biserial correlation coefficient arises. In biserial correlation, Y and X' are normally distributed random variables and their bivariate distribution is such that Y has a linear regression on X'. Suppose in the bivariate population X' is dichotomized. A random sample is drawn in which observations are made on Y and on the dichotomization of X'. It is desired to estimate $\rho_{x'y}$, the population Pearson product-moment correlation coefficient, from the sample. Karl Pearson (1909) proposed a sample estimator of $\rho_{x'y}$ which has been adopted almost exclusively. It is well known that $\rho_{x'y} = \beta_{x'y}$ $\sigma_{x'}/\sigma_y$, where $\beta_{x'y}$ is the slope in the population of the least-squares regression line of Y on X' and σ denotes the standard deviation of a variable. The standard deviation of Y can easily be estimated from the sample, and Pearson showed how $\sigma_{x'}$, could be estimated from the dichotomously scored sample. The slope of the regression line, $\beta_{x'y}$, was estimated by the slope of the line passing through the two points (0, $\bar{Y}_0$) and (1, $\bar{Y}_1$), where $\bar{Y}_0$ and $\bar{Y}_1$ are the means on Y of those scoring 0 and 1 on the dichotomy, respectively. This line is such that the sum of the squared distances along the Y-axis of the sample points from the line is minimal (see Kelley, 1947, pp. 373–4). Substituting the slope of this sample regression line and estimates of $\sigma_{x'}$ and σ_y for the corresponding parameters in $\beta_{x'y}$ $\sigma_{x'}/\sigma_y$ yields the conventional biserial correlation coefficient:

$$r_{\text{bis}} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_y} \frac{pq}{z},$$

where

s_y^2 is the unbiased estimator of σ_y^2,
$p = 1 - q$ is the proportion of 1's on the dichotomy,
z is the ordinate on the unit normal curve above the $(100p)$th percentile of the distribution.

The coefficient r_{bis} is an estimator of $\rho_{x'y}$, the population product moment correlation coefficient between the two undichotomized variables. Tate (1955) showed that r_{bis} is a consistent estimater of $\rho_{x'y}$;

thus in the bivariate normal population the product-moment and the biserial correlation coefficients are equal.

Suppose now that Y and X' are ranking variables, and that X is a dichotomous variable derived from X' in the manner indicated in the second paragraph in this section. In the population of individuals, the product-moment correlation of the ranking variables Y and X' is the population Spearman rank-order correlation coefficient. From a sample of n persons for which observations on Y and the dichotomy X are available, we seek an estimate of this population coefficient. An estimate can be derived in the same manner that r_{bis} is derived as an estimate of ρ. The desired estimator of the population Spearman's rho from a set of n ranks (observations on Y) and n dichotomous observations, X, will be denoted rb. The sample value of rb will have the form:

$$rb = b_{yx'} \frac{s_{x'}}{s_y},$$

where

$b_{yx'}$ is an estimate of the least-squares regression coefficient of Y on X',
$s_{x'}$ is the standard deviation of X', and
s_y is the standard deviation of Y.

It is well known that $s_{x'}^2 = s_y^2 = (n^2 - 1)/12$ (see Siegel, 1956, p. 203). Hence, $rb = b_{yx'}$. Following the estimation procedure used in deriving the classical biserial coefficient, we let $b_{yx'}$ equal the slope of the line which passes through the two points $(\bar{X}_0', \bar{Y}_0)$ and $(\bar{X}_1', \bar{Y}_1)$, where the bar above the variable denotes an average and the subscript to the variable denotes one group of the dichotomy. For example,

$\bar{Y}_0 = \sum_1^{n_0} Y_{i0}/n_0$, where the subscript i ranges over persons.

The slope of the chosen line is the coefficient rb and is given by the equation

$$rb = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1' - \bar{X}_0'}.$$

$\bar{Y}_1$ and $\bar{Y}_0$ are easily calculable from the observed ranks, Y. $\bar{X}_1'$ and $\bar{X}_0'$ are calculated as follows:

$$\bar{X}_0' = \sum_1^{n_0} X_{i0}'/n_0.$$

$$\sum X_{i0}' = 1 + 2 + \cdots + n_0 = n_0(n_0 + 1)/2.$$

Therefore, $\bar{X}_0' = (n_0 + 1)/2$.

$$\sum X_{i1}' = (n_0 + 1) + (n_0 + 2) + \cdots + (n_0 + n_1) = n_1 n_0 + n_1(n_1 + 1)/2.$$

Therefore, $\bar{X}_1' = n_0 + (n_1 + 1)/2$.
Thus, $\bar{X}_1' - \bar{X}_0' = n/2$.

It follows, then, that rb has the form

$$rb = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1' - \bar{X}_0'} = \frac{2}{n}(\bar{Y}_1 - \bar{Y}_0). \tag{1}$$

The coefficient rb has an attractively simple definitional form. Computational formulas for rb can be found which are even simpler. These computational formulas follow:

$$rb = \frac{2}{n_0}\left[\bar{Y}_1 - \frac{n+1}{2}\right], \tag{2}$$

$$rb = \frac{2}{n_1}\left[\frac{n+1}{2} - \bar{Y}_0\right]. \tag{3}$$

Formulas (2) and (3) are equivalent. One of the two will probably be very much simpler, depending on the sizes of n_1 and n_0 and the size of the ranks in one group of the dichotomy. (Note that rb is not defined when either n_1 or n_0 is zero.) The computation of rb will be illustrated on the data below:

The ranks of those scoring *1* on the dichotomy appear under column *1*.

Scores on Y for $X = 1$	$X = 0$
10	8
9	6
7	5
4	3
	2
	1

Calculations

$n_1 = 4,\ n_0 = 6$
$\bar{Y}_1 = 30/4$
$rb = (2/6)(30/4 - 11/2) = 2/3.$

The significance of the difference of rb from zero can be tested by an independent ranks test such as the Mann-Whitney U-test. One

simply runs the ranks test on the ranks of those scoring *1* on the dichotomy against those scoring *0* on the dichotomy. (See Siegel, 1956.)

Cureton's Rank-Biserial Coefficient

Cureton (1956) developed a coefficient descriptive of the correlation between a ranking variable and dichotomous variable. We shall denote his coefficient by r_c for the time being. As compared with the coefficient developed in the previous section, r_c makes no assumption of a ranking variable underlying the dichotomy, and it is calculated by counting inversions and agreements; hence it is not a product-moment coefficient. The methods of computations for rb and r_c are quite dissimilar. Oddly enough for untied ranks rb and r_c always give the same value for a set of scores. This will be proved in the following section. In this section the derivation of r_c will be outlined.

Let X be a dichotomous variable and Y a variable comprising the n untied ranks 1, 2, . . ., n. Cureton sought a coefficient descriptive of the relationship between X and Y such that (a) it would have attainable limits ± 1 under all circumstances, (b) it would be $+1$ when the n_1 highest ranks are all *1* on the dichotomy, (c) and it would be strictly nonparametric, i.e., defined wholly in terms of inversions and agreements without such concepts as mean, variance, regression, etc.

To compute r_c, the data are arranged in the following manner:

Scores on Y for $X = 1$	$X = 0$	*Agreements*	*Inversions*
10		6	
9		6	
	8		2
7		5	
	6		1
	5		1
4		3	
	3		
	2		
	1		
		$P = 20$	$Q = 4$

There is an *agreement* at any given rank under column *1* for every smaller rank under column *0*. There is an *inversion* at any given rank under column *0* for every smaller rank under column *1*. Thus there are three agreements corresponding to the rank "4" under column *1* since there are three smaller ranks, 3, 2, and 1, under column *0*. P is the sum of all agreements in the data, and Q is the sum of all inversions.

Cureton defined r_c as follows:

$$r_c = (P - Q)/P_{\max}. \tag{4}$$

When no ties exist (the only case we have considered throughout this paper) $P_{\max} = n_0 n_1$; again, n_0 is the number of persons at 0 on the dichotomy and n_1 is the number of persons at *1*. Hence

$$r_c = (P - Q)/n_0 n_1. \tag{5}$$

For the above data,

$$r_c = (20 - 4)/4 \cdot 6 = 16/24 = 2/3.$$

Notice that rb and r_c have the same value, 2/3, for the above data. In the next section it will be shown that rb and r_c are algebraically equivalent in the no ties case, even though they are quite different in their inception, derivation, and computation. The proof to follow will point out similarities between the process of counting agreements and inversions and the arithmetic manipulation of ranks.

Proof that $(P - Q)/(n_0 n_1) = (2/n)\ (Y_1 - Y_0)$, *i.e.*, $r_c = rb$

It will be convenient to express $(2/n)\ (Y_1 - Y_0)$ in a different form.

$$\frac{2}{n}(\bar{Y}_1 - \bar{Y}_0) = \frac{2}{n_0}\left[\bar{Y}_1 - \frac{n+1}{2}\right] = \frac{2}{n_0 n_1}\left[\sum Y_1 - \frac{n_1(n+1)}{2}\right].$$

To establish the proof it is sufficient to show that

$$P - Q = 2\left[\sum Y_1 - \frac{n_1(n+1)}{2}\right].$$

We note that $P + Q = n_1 n_0$, hence $P - Q = 2P - n_1 n_0$. Therefore, we shall prove that

$$\begin{aligned} 2P - n_1 n_0 &= 2 \sum Y_1 - n_1(n+1), \text{ or} \\ 2P &= 2 \sum Y_1 - n_1(n+1) + n_1 n_0, \text{ or} \\ P &= \sum Y_1 - [n_1(n_1+1)]/2 \end{aligned} \tag{6}$$

Establishing (6) will establish that $r_c = rb$.

The following example will illustrate the proof:

1	*0*	*Agreements*	*Inversions*
6		3	
	5		2
4		2	
3		2	
	2		
	1		
		$P = 7$	$Q = 2$

We shall show that a method of counting and summing agreements is equivalent to evaluating $\Sigma Y_1 - [n_1(n_1 + 1)]/2$.

"Agreements" are defined only for those ranks which appear under column *1*. We shall always arrange the ranks from highest to lowest. Suppose we wish to find the number of agreements associated with the smallest rank under column *1*, 3 in the above example. Denote this smallest rank under column *1* by Y_{11}. Now Y_{11} scores an agreement every time it exceeds a smaller rank under column *0*. Obviously, Y_{11} will always exceed $Y_{11} - 1$ ranks which appear under column *0*. How many agreements are associated with Y_{21}, the next to the smallest rank under column *1* (4 in our example)? There are only $Y_{21} - 1$ ranks that are smaller than Y_{21} and all but one of these is in column *0* (Y_{11} is in column *1*); hence there are $Y_{21} - 2$ agreements associated with Y_{21}. How many agreements are associated with Y_{31}? Of the $Y_{31} - 1$ ranks smaller than Y_{31}, two of them, Y_{21} and Y_{11}, are in column *1*); hence, there are $Y_{31} - 3$ agreements associated with Y_{31}.

It should be easy to see that in general there are $Y_{j1} - j$ agreements associated with the jth smallest rank under column *1*. The sum of these agreements is P, which we find as follows:

$$P = \sum_{1}^{n_1} (Y_{j1} - j) = \sum_{1}^{n_1} Y_1 - [n_1(n_1 + 1)]/2.$$

This establishes the equivalence of r_c and rb.

Discussion

It has been shown that a *tau*-type coefficient, which is not a correlation coefficient at all, derived by Cureton (1956), for use with a dichotomous and a ranking variable is algebraically equivalent to

a coefficient derived in this paper which estimates Spearman's *rho* between the ranking variable and a hypothetical ranking variable assumed to underlie the dichotomy. The latter coefficient is clearly of the nature of a product-moment correlation coefficient. The equivalence holds only for the case of no ties in the observed ranking variable. When ties occur, *rb* and r_e are no longer equivalent.

These strange equivalences cropping up in various places in statistics seem to indicate basic similarities between apparently different concepts. For example, it is not widely known that Spearman's *rho* is the same simple function of *weighted* agreements and inversions that Kendall's *tau* is of *unweighted* agreements and inversions (Durbin and Stuart, 1951).

Of what possible use is *rb*? (1) It measures monotonicity of relationship between the variables underlying Y and X', (2) it always lies on the interval -1 to $+1$, inclusive, (3) it is computationally simple. These points are elaborated below:

1. Unlike a correlation ratio which looks for any possible relationship between two variables and unlike Pearson's r which measures *linearity* of relationship between two variables, *rb* measures an intermediate type of relationship between the variables underlying the two sets of ranks. The value of *rb* is a measure of the monotonicity of relationship between the variables underlying the two sets of ranks, *Y and X'*. *Monotonicity* is more general than *linearity* and less general than merely *any sort of relationship.*
2. If we exclude those instances of no variance in either Y or X (which all correlation measures do), then it can be shown that *rb* has a minimum of -1 and a maximum of $+1$ regardless of the extremeness of the dichotomization on X or the nature of the variables underlying Y and X'. Researchers often experience the unpleasantness of obtaining Pearson biserial correlation coefficients with values as high as 1.20.
3. Formulas (2) and (3) above are particularly simple, and either (2) or (3) will be simpler depending on a given set of data. For untied ranks, formulas (2) and (3) give Cureton's coefficient without listing data and counting inversions and agreements.

REFERENCES

Cureton, E. E. Rank-biserial Correlation. *Psychometrika,* 1956, 21, 287–290.

Durbin, J. and Stuart, A. Inversions and Rank Correlation Coefficients. *Journal of the Royal Statistical Society,* 1951, 13, 303–309.

Kelley, T. L. *Fundamentals of Statistics.* Cambridge, Massachusetts: Harvard University Press, 1947.

Pearson, K. On a New Method for Determining the Correlation between a Measured Character A, and a Character B. *Biometrika,* 1909, 7, 96–105.

Siegel, S. *Nonparametric Methods for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

Tate, R. F. The Theory of Correlation between Two Continuous Variables When One is Dichotomized. *Biometrika,* 1955, 42, 205–216.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 633-641.

ESTIMATING GAINS IN SCHOLASTIC APTITUDE TEST SCORES ATTRIBUTABLE TO THREE SOURCES

SAM C. WEBB
Georgia Institute of Technology

IN recent years scores on the Scholastic Aptitude Test (SAT) have become increasingly important as measures of the intellectual quality of entering college freshman classes. Class profiles showing distributions of scores for applicants and for accepted and enrolled students are routinely made available as a part of publicity and recruiting materials. Conversations concerning the quality of college classes tend to focus on average SAT scores as the principal index of academic ability. Many institutions, desirous of improving their image in the academic world, point to gains in scores as evidence that admission standards have been raised. And changes in average scores have often been regarded as evidence for the increased effectiveness of recruiting and selection procedures.

Though the high school average (HSA) or some derivative such as rank in class is for most colleges a more valid predictor of college performance than are SAT scores, the former seems to receive less emphasis in these kinds of evaluations. There are understandable reasons why this is so.

SAT scores, for instance, are more objective than the HSA in that the SAT applies the same scale of measurement to all students regardless of the high school from which they come. High school averages, on the other hand, vary from school to school so that values of different schools are often not comparable; and class averages computed from records of different schools are in a technical sense somewhat meaningless. More importantly, perhaps, the very nature of the grade scale as usually employed tends to obscure differences between grading standards among schools and

changes in standards within schools. The result is, of course, that the HSA is a rather poor measure for assessing changes in the quality of admitted college classes.

As already intimated, there have been in recent years gains in average SAT scores for entering freshman classes at a number of colleges. These gains usually have been interpreted as signifying a rise in the intellectual quality or grade-getting abilities of students. This sort of interpretation, however, has not received unanimous support. Some professors seem to feel that because of the relatively small increments in HSA averages and because of impressions they have over the years of grades made by students in their classes, these gains in average scores must be statistical artifacts emanating from the procedures employed in scaling the scores. However, since there are technical studies to show the scoring scales have remained quite stable over a considerable period of time (Angoff and Waite, 1959, 1961) and since there are empirical data that show practically no change in average SAT scores for all college applicants as reported by the College Entrance Examination Board from 1957 to 1960 (Fishman, 1957, 1962), this kind of reasoning does not seem to provide an adequate explanation for the observed gains.

Thus other explanations must be sought. Reasonably cogent and plausible explanations come to mind which suggest that the causes underlying reported gains may well be numerous and complex.

These suggested causes may be grouped into two broad categories: those that cannot be attributed directly, and perhaps not even indirectly, to the activity or influence of any particular college or colleges; and those that are essentially a function of the nature and activities of a college (or colleges) or closely associated individuals (or groups of persons). For want of better terminology these causes may be referred to as "non-college" and "college" causes.

Several causes may be classified in the "non-college" group. These include a variety of factors which are psychological, educational, sociological, psychometric, and perhaps even physiological in character. Some possibilities that readily come to mind are improved instruction and preparation for college in high school, short term coaching effects—especially on the mathematical section for students not taking math courses—increased test-taking ability or test "wiseness," practice effects, and increased societal incentives

(psychological, sociological, and economic) which encourage increased numbers of able students to apply for admission to college. Probably of lesser importance but also worthy of inclusion here is the possibility of increased basic intellectual capacities of the population over time as a function of nutritional and physiological factors.

The second or "college" factors category may be divided into two classes. In the first are included all those factors which are influential in causing a student to apply for admission at a particular college. For convenience these are called "recruitment" factors. Suggested for inclusion in this class are various characteristics of the college—both actual and imagined; family associations and ties with the college; persuasive activities of friends, acquaintances, and peers; and the publicity and recruitment activities of the college.

In the second class are included those factors which are effective in the student's enrollment in a particular college. Here are grouped all those factors which cause a student to be admitted by the college and which cause the student to enroll. For convenience these are called "admission and acceptance" factors. Associated with the selection process are the admissions procedures and policies of the college and the characteristics of the student and his credentials and associations which are influential in his being selected for admission. Associated with the acceptance process are such factors as the scholarship or other financial resources available to the student through the college; the image of the college; and influences of family, friends, acquaintances and friends already mentioned.

Except for the effects of short term coaching programs, there seems to be a scarcity of published studies on the effects of any of the suggested causative factors. This is regretable, for it would be both interesting and useful to assess what influence each of these suggested factors has in determining changes in average SAT scores. With such knowledge individual colleges could more meaningfully evaluate the effectiveness of their recruitment and selection procedures. And on a larger scale, perhaps, the relative importance of these factors as they affect students differentially among colleges could be investigated.

Unfortunately, however, the very variety of possible causative

factors, and the probably high inter-relationships among some of them, leads one to suspect that rather complicated statistical and experimental designs and some rather expensive and long-term research would be required to secure satisfactory estimates of the effects they are producing. Consequently precise answers as to the effects of all these factors will probably not be forthcoming in the near future.

In the meantime some approximate but nevertheless useful estimates of the effects of the three groups of factors already described—non-college, recruitment, and selection and acceptance—can be easily made when certain data are available.

Three sets of average SAT scores, preferably broken down by test section and sex, are required. Each set should include averages for the beginning and ending of the same arbitrarily selected period of time. The first set should provide valid and unbiased estimates of the average scores for a base population comprised of high school students who will seek admission to college and who reside in a defined geographical area. Preferably this should be an area from which the college draws its students. The second set of data should provide average scores for the applicant population of the college; and the third set should provide average scores for applicants who enrolled as freshmen.

From these data an estimate of the effects of "non-college" causes on score changes over the period can be obtained by using averages for the base population and by subtracting averages for the beginning of the period from those for the end of the period. Estimates of the effects of the "recruitment" factors can be obtained in two stages. An estimate of the effects of these factors at the beginning of the period may be obtained by substracting average scores for the base population at the beginning of the period from average scores for the college applicant population at the beginning of the period. Gain (or loss) in scores over the period resulting from recruitment factors may be obtained by first getting the difference between the applicant population and the base population average at the end of the period, and subtracting from this value the difference between the applicant population and base population average at the beginning of the period.

In like fashion changes attributable to selection and acceptance factors can be estimated in two stages. Subtracting average scores

for students enrolled from averages for the college applicant population at the beginning of the period provides an estimate of selection effects obtaining at the beginning of the period. Gains resulting from these effects during the period may then be obtained by getting the difference between averages for enrolled students and the college applicant population at the end of the period and subtracting from this value gains existing at the beginning of the period already obtained.

From these values total changes over the period resulting from all causes can be estimated. It is also possible to determine the total difference between the quality of students enrolled at the end of the period as contrasted with the regional population at the beginning of the period.

The use of these procedures is illustrated with data for Emory College and for all predominately white public colleges of the state of Georgia (Klock, 1964) for the five year period covering the academic years of 1958-59 through the year 1962–63. The SAT averages employed are shown in Tables 1 and 2. In making estimates it

TABLE 1

Average SAT Scores for Applicant Populations of Emory College and Georgia White Public Colleges

Academic Year	Emory College				Georgia Colleges			
	SAT-V		SAT-M		SAT-V		SAT-M	
	Male	Female	Male	Female	Male	Female	Male	Fem
1958–59	475 ($n = 717$)	495 ($n = 586$)	509	476	400 ($n = 4013$)	387 ($n = 2068$)	457	38
1962–63	519 ($n = 1094$)	528 ($n = 931$)	558	527	421 ($n = 5796$)	421 ($n = 3281$)	472	41

TABLE 2

Average SAT Scores for Students Entering Emory College and Georgia White Public Colleges

Academic Year	Emory College				Georgia Colleges			
	SAT-V		SAT-M		SAT-V		SAT-M	
	Male	Female	Male	Female	Male	Female	Male	Fe
1958–59	504 ($n = 275$)	505 ($n = 195$)	534	492	400 ($n = 4013$)	387 ($n = 2068$)	457	3
1962–63	548 ($n = 368$)	570 ($n = 240$)	589	575	439 ($n = 4201$)	432 ($n = 2397$)	494	4

has been assumed that the applicant population values for the Georgia system are valid and unbiased estimates of the base population of the region from which Emory College and the Georgia system draw their applicants. Associated with this assumption is the further one that recruitment effects for the system as a whole are of minimal consequence. It is further assumed that the data for the enrolled students for the Georgia public colleges in 1958–59 are a valid estimate of the applicant population of the system for that year. This assumption seems justified on the basis of information which suggests that, except for the requirement of graduation from high school, there was practically no selectivity in admissions for the system as a whole in that year.

The estimates for Emory College are given in Table 3; those for all Georgia white public colleges are given in Table 4. The large number of exact zero entries in Table 4 appear because 1958–59 averages for enrolled students of the Georgia system have been used as estimates for the base population and for the applicant population for the Georgia system for that year.

TABLE 3

An Analysis of Gains in SAT Scores for Emory College from 1958 through 1962

Sources of Gain	SAT-V Male	SAT-V Female	SAT-M Male	SAT-M Female	Unweighted Average
Non-College Factors	+21	+34	+15	+32	+25.5
Recruitment Factors:					
Gain at beginning of period	75	108	52	90	81.3
Gain at end of period	98	107	86	109	100.0
Gain over period	+23	−1	+34	+19	+18.7
Selection and Acceptance Factors:					
Gain at beginning of period	29	10	25	16	20.0
Gain at end of period	29	42	31	48	37.5
Gain over period	0	+32	+6	+32	+17.5
Total Gain over 5 years	44	65	55	83	61.8
Total Gain of 1962 enrolled students over 1958 base population	148	183	132	189	163.0

TABLE 4

An Analysis of Gains in SAT Scores for All Georgia White Public Colleges from 1958 through 1962

Sources of Gain	SAT-V Male	SAT-V Female	SAT-M Male	SAT-M Female	Unweighted Average
Non-College Factors	+21	+34	+15	+32	+25.5
Recruitment Factors:					
Gain at beginning of period	0	0	0	0	0
Gain at end of period	0	0	0	0	0
Gain over period	0	0	0	0	0
Selection and Acceptance Factors:					
Gain at beginning of period	0	0	0	0	0
Gain at end of period	18	11	22	9	15.0
Gain over period	+18	+11	+22	+9	+15.0
Total Gain over 5 years	39	45	37	41	40.5
Total Gain of 1962 enrolled students over 1958 base population	39	45	37	41	40.5

For Emory College the data show unweighted average gains taken across tests and sexes during the period of about 26 points arising from non-college factors, of about 19 points from recruitment factors and of about 18 points from selection and acceptance factors. This yields an average gain from all causes of about 62 points over the five year period. For white public colleges of Georgia there was the same 26 point gain from non-college factors and a 15 point gain from selection and acceptance factors. Thus there is about a 41 point gain over the period due to all factors. This means that there is about a 21 point greater gain for Emory than for the Georgia colleges.

While this discussion has emphasized the analysis of gains within the five year period, it is clear from Tables 3 and 4 that the analysis also yields information concerning the magnitude of gains over the base population attributable to recruitment and to selection and acceptance factors existing at the beginning of the period. Thus for Emory College at the beginning of the period there were aver-

age gains of 81 and 20 points attributable to these two types of sources respectively. Further it can be noted that the average gain for the 1962 enrolled students over the 1958 base population was 163 points for Emory College and 40 points for all Georgia white public colleges.

For Emory College it is also interesting to note that while differences from recruitment factors in 1958–59 were larger for females than for males, the gains over the five year period were larger for males than for females. Also while gains attributable to selection and acceptance factors were greater for boys than for females at the beginning of the period, gains over the period were considerably larger for females. This latter difference is largely a function of limited housing facilities for women. In the aggregate, gains from both recruitment and selection and acceptance factors are larger for females than for males.

In making estimates of the type described here, the greatest problem is that of finding averages that will be accepted as valid and unbiased estimates for an appropriately selected base population. In the illustrations used, for example, had data for the applicant and enrolled populations for Emory and the Georgia system for the years 1957–58 and 1960–61 been available, all applicants in the country taking the SAT for those years could have been used as the base population, since the College Entrance Examination Board has published average scores for this group for those years. However, since Emory draws the majority of its students from the Southeast, and since the Georgia system draws its students primarily from the state of Georgia, a base population of college applicants in this region would seem more appropriate than one involving a larger geographical area, even though the estimates of averages may be less accurate.

It should be noted that only three estimates are affected by the data for the base population. These are the estimates of "non-college" effects, and the two estimates of recruitment effects. And one may judge the effects that base population values considered to exceed or fall below some "true" value have on these estimates. For example, if one considers both base population values to be overestimated to the same degree, the estimate of gains due to recruitment factors at the beginning of the period will be underestimated. Then suppose one considers the base population values at

the beginning of the period to be underestimations and those at the end to be about right or over-estimations. Then gains due to non-college factors will be over-estimated; gains due to recruitment factors at the beginning of the period will be over-estimated; and gains due to recruitment factors over the period will be under-estimated.

In respect to the illustrative data used, several inquiries lead the writer to conclude that the data for the base populations are the best that are presently available for Georgia and the Southeastern region. If these values deviate from the "true" population values, they probably do so in the direction of being underestimates as a function of the abler students enrolling in private colleges or colleges of other state systems. For economic reasons, these deviations are probably larger for the 1962–63 data than for the 1958–59 data. If these speculations are correct, then gains attributable to non-college factors are underestimated and gains attributable to recruitment factors both at the beginning of the period and over the period are overestimated.

In conclusion, it is to be emphasized that the procedures described provide a suggested way for estimating the contributions of three types of causes of changes in average SAT scores that occur over a period of time for enrolling freshman classes in a college or group of colleges. While the illustrative data involved only gains, the procedures seem general in application and could apply to any type of change. The results seem potentially useful in getting a reasonable evaluation of the influence of these groups of factors on changing average SAT scores for a college.

REFERENCES

Angoff, W. H. and Waite, A. C. A Study of Double Part-score Equating for the Scholastic Aptitude Test. SR–59–29. Educational Testing Service, Princeton, N. J., August, 1959. Anon "'Freeze Present College Board Scale, Use Diverse Norms,' Wilks Study Bids." *ETS Developments* No. 1, Vol. X, Educational Testing Service, Princeton, N. J., November 1961.

Fishman, J. A. *1957 Supplement to College Board Scores, No. 2.* New York: College Entrance Examination Board, 1957. College Entrance Examination Board, *College Board Score Reports: A Guide for Counselors.* Princeton, N. J., 1962.

Klock, J. A. An Investigation of Increases in Mean SAT Scores. *Research Bulletin* 64–6, Office of Testing and Guidance, Board of Regents University System of Georgia, September 1964.

Educational and Psychological Measurement
1966, 26, 643-659.

A STUDY OF SAMPLE SIZE IN MAKING DECISIONS ABOUT INSTRUCTIONAL MATERIALS[1]

LAWRENCE M. STOLUROW
Harvard University
AND
GERALD FRINCKE
Sacramento State College

The present study was conducted to explicate the problems involved in making decisions about the acceptability of frames in self-instructional programs. Its approach, different from that of Holland (1965) using the "blacked-out" procedure to determine the effects of stimulus degradation on error rate and that suggested by Stolurow (1965) using the redundancy test. In the present study, data were examined to determine the relative efficiency of various sample sizes in making decisions about retaining or revising the frames of a self-instructional program. Individual frames are not assumed to be statistically independent of one another. The errors made on one frame are likely to be correlated with those made on another and the magnitude of the correlations is generally both unknown and variable throughout a program. This state of affairs indicates the need for an empirical study of sampling problems in decision making. Two indices were used to determine the implications of selection criteria and sampling procedures in deciding about the acceptability of frames. One index was the per

[1] This paper is based upon research supported, in part, by the U. S. Office of Education, Educational Media Branch, Title VII, Grant No. 7-23-1020-151.1. Co-investigators were Lawrence M. Stolurow and Max Beberman. This study was issued originally as Technical Report No. 11, February, 1965, by the Training Research Laboratory, University of Illinois, under the same authorship and title. Reproduction in whole or in part is permitted for any purpose of the United States Government.

cent of undesirable frames correctly identified. The other was the per cent of rejections made erroneously (Type I errors). The study also was conducted to suggest guidelines for determining the sample size to use in developing self-instructional programs in which it is expected that the overall error rates will be low and the distribution of error rates observed for frames within the program will be skewed.

Decision Making and Types of Errors

In the course of evaluating and revising programed instructional materials, a programer may decide to reject as undesirable those frames in the program which he suspects will lead to a student error rate above a given criterion value. In doing this, the programer is in a position similar to that of a statistician faced with a large number of hypotheses to test. Each frame in the program must be accepted or rejected. In effect, for each frame, the programer must test the null hypothesis that the frame will have an error rate below that value he has chosen as his minimum criterion for "undesirableness" when the program is put into general use. The number of observations on which each test of this hypothesis is based will be equal to the number of students with whom the program is pretested. Usually (though not necessarily) the programer rejects the null hypothesis whenever a frame is found to have an error rate equal to or greater than the one chosen as the criterion value.

In making each decision concerning the acceptability of a program frame, the programer, as statistician, may make one of two types of errors. He will make a Type I error if the null hypothesis is rejected when it is true; he will mislabel an acceptable frame as "unacceptable." He will make a Type II error if the null hypothesis is accepted when it actually is false; he will mislabel an unacceptable frame as "acceptable." The programer definitely wants to avoid Type II errors. He wants what the statistician calls a "powerful" test of the null hypothesis, one that has enough "power" to reject each truly unacceptable frame. At the same time, however, the programer does not want to make too many Type I errors. Otherwise, he will be needlessly rewriting a large number of the acceptable frames in the program. Extensive rewriting would then require that the revised program be pretested, and thus delay and increase the cost of the finished program.

Power, Sample Size and Student Ability

Given a fixed number of observations (N), the only way to change the probability of rejecting an acceptable frame is to shift the rejection criterion. Unfortunately, in this situation a shift of the rejection criterion which reduces chances of this Type I error simultaneously reduces the power of the test. Conversely, a shift of the criterion which increases the power of the test results in the increased probability of a Type I error. Only by increasing N can the power of the test be increased without a simultaneous increase in the probability of a Type I error. Similarly, only by increasing N can the probability of a Type I error be reduced without diminishing the power of the test.

Since each test is based on the error data of students utilized in pretesting the program, the power of each test will be inversely related to the performance level or overall error rate of the N students comprising the sample. If the N students perform quite well, error rate estimates will be depressed for all items. Consequently, fewer unacceptable items will be rejected. Conversely, if the N students perform quite poorly, observed error rates will be higher and more frames (both acceptable and not) will be rejected. This suggests the desirability of obtaining measures characterizing the sample of students used in terms of relevant abilities, and of obtaining measures of their representativeness in terms of relevant academic achievement.

True Error Rate and Rejection

We can speak of the "true" error rate of a frame as that error rate which would be found if the frame were given to all of the intended population as a part of the finished program. The power of each test made by the programer also depends upon the degree to which the actual state of affairs approaches that stated in the null hypothesis. Thus "unacceptable" frames with very high true error rates are more likely to be rejected on the basis of pretesting data than unacceptable frames with true error rates which more closely approach acceptability. The programmer may thus expect that the percentage of truly unacceptable frames rejected by tests based on data from a given pretesting sample of size N will not

be very high in the case of programs which have already undergone fairly successful revision.

Method

Materials

The program. The University of Illinois Committee on School Mathematics, (UICSM) programed instructional series, Part 110, was selected for study since it teaches relatively difficult mathematical concepts and thus some variability in error rates for frames could be expected. Worksheet error data for 178 of the students who completed (pure mode) Part 110 (Beberman and Stolurow, 1963) were recorded on SCRIBE[2] sheets and used with the SCRIBE system to produce IBM cards[3] containing individual error data for each student. Error data from the completed worksheets of the 52 additional pure-mode subjects were punched directly into IBM cards in the SCRIBE output format. Thus, data from 230 students were available for sampling.

Sampling procedure. All sampling of data from the basic pool of 230 students was done without replacement. A stratified random sample of 100 students from all seven classes in four participating schools was selected to establish criterion error rate measures on the 308 program frames in UICSM PIP Part 110. A similar stratified random sampling procedure was then followed as closely as possible in selecting several other samples of various sizes. In cases where a stratified procedure was clearly impossible, such as in a sample of size 1 ($N = 1$), ordinary random sampling was utilized. Three samples each of the following sizes were selected: $N = 1$, $N = 2$, $N = 3$, $N = 4$, $N = 5$, $N = 10$, and $N = 15$. These samples were also merged to form a "summation sample" with $N = 120$.

---

[2] SCRIBE is a system developed and used by ETS to score multiple-choice answer sheets and to transcribe automatically the data on the answer sheets to IBM cards.

[3] The SCRIBE technique of recording and processing worksheet data is described by Frincke, G. L. and Stolurow, L. M. Three methods of recording worksheet performance. Urbana, Illinois: University of Illinois, 1964. USOE Title VII, Technical Report No. 7. This work was done in cooperation with Educational Testing Service and arrangements for it were made by Dr. Paul Jacobs. The ETS contribution was made possible by a grant from the Carnegie Corporation of New York.

Item analyses.[4] Worksheet error data cards prepared from the worksheets of the subjects who constituted the criterion sample, the summation sample, and the 21 smaller samples, were used as the basis for 23 separate item-analyses of the 308 items which comprise the UICSM Programed Instruction Part 110.

Three hundred and eight summary IBM cards were then prepared. Each of these cards contained the results of all 23 analyses with regard to one of the 308 program items. The summary cards were used to determine correlations between the results of item analyses based on the various samples. These cards were also used to determine which frames would be rejected and which accepted by tests based on the data of the various samples in cases where the criterion for rejection would be the observed error rates equal to or greater than 10%, 15%, or 20%.

Results

Figure 1 is a frequency distribution of the different error rates observed in Part 110 for the criterion sample ($N = 100$). Table 1

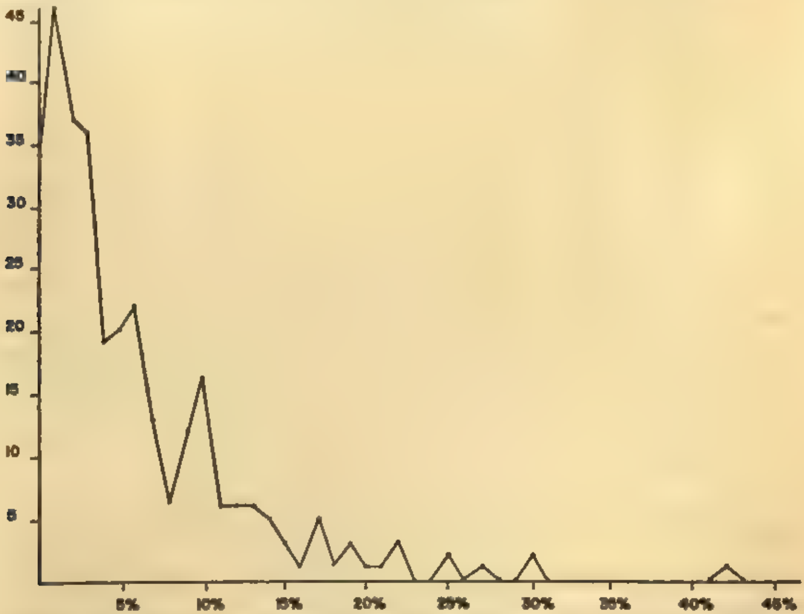


Figure 1. Frequency distribution of error rates as determined with the criterion sample ($N = 100$).

[4] These analyses were carried out with the aid of an IBM 1620 computer and a program written by Scott Krueger, University of Illinois, Training Research Laboratory.

shows the distribution of error rates observed in the 21 smaller samples. All of the distributions are quite skewed. Most items are well within the limits of acceptability and the number of extremely unacceptable items in Part 110 is actually quite low. This is an important factor in interpreting the findings of this study.

TABLE 1

Distribution Errors in Each in 21 Small Samples[a]

Sample size	Number of students who made errors									Total frames
	0	1	2	3	4	5	6	7	8–15	
1	286	22								308
1	300	08								308
1	296	12								308
2	261	45	2							308
2	266	41	1							308
2	287	20	1							308
3	267	37	3	1						308
3	271	32	5	0						308
3	257	50	1	0						308
4	288	15	5	0	0					308
4	216	79	11	2	0					308
4	218	80	8	1	1					308
5	177	95	25	9	2	0				308
5	220	69	15	4	0	0				308
5	241	61	5	1	0	0				308
10	225	56	21	3	1	2	0	0	0	308
10	152	100	38	8	5	5	0	0	0	308
10	190	79	30	5	4	0	0	0	0	308
15	194	76	26	9	0	1	1	1	0	308
15	186	75	27	8	8	2	1	1	0	308
15	159	89	47	5	5	2	0	1	0	308

[a] Based on UICSM Programed Intruction Booklet Part 110.

Table 2 presents correlations between the error rates of the 308 items as determined by each of the samples and error rates based on the criterion sample. These correlations are generally quite low as might be expected due to the small range of error rates obtained and the extreme skewness of the error rate distributions.

Estimates of the overall program error rate based on the 21 independent samples also are presented in Table 2. Considerable variability among the estimates based on the smaller samples can be seen. For example, those for samples of size 1 range from 2.5 per cent to 7.1 per cent, those of size 5 range from 4.8 per cent to 11.7 per cent overall error rate. Figure 2 shows the mean overall error rate estimates for each of the sample sizes in relation to the over-

TABLE 2

Correlations of Sample Item Analyses with Criterion Item Analyses as a Function of Sample Size

Sample		Estimate of program error rate in %	Correlation with criterion[c]
Code No.	Size		
0[a]	100	5.6	1.000
1	1	7.1	.329
2	1	2.5	.288
3	1	3.6	.101
4	2	8.0	.356
5	2	6.9	.297
6	2	3.6	.318
7	3	5.0	.272
8	3	4.5	.352
9	3	5.6	.302
10	4	2.0	.455
11	4	8.7	.481
12	4	8.4	.462
13	5	11.7	.453
14	5	7.2	.470
15	5	4.8	.535
16	10	3.9	.567
17	10	8.0	.514
18	10	5.5	.552
19	15	3.7	.680
20	15	4.5	.710
21	15	5.1	.641
22[b]	120	5.5	.872

[a] Criterion sample.
[b] Summation sample.
[c] Analysis is based upon 308 frames of UICSM PIP Part 110. All correlations except .101 are significant beyond the .01 level.

all error rate observed in the criterion sample. All of these overall rates appear to fluctuate randomly about the criterion value with the exception of that based on the samples of $N = 5$. Here a very excessive and presumably atypical error rate was observed.

Figure 3 depicts the efficiency of the seven sample sizes by showing the relationship between the mean percentage of unacceptable frames that were rejected and sample size. It shows that all sample sizes greater than $N = 4$, with the exception of the sample size $N = 10$ at the 15 per cent criterion level, rejected 50 per cent or more of the unacceptable items. In all cases, a rapid rise in the mean percentages of the unacceptable items correctly identified for rejection can be seen in relation to the sample size as it increases from one to five. Each line represents a different criterion. In the

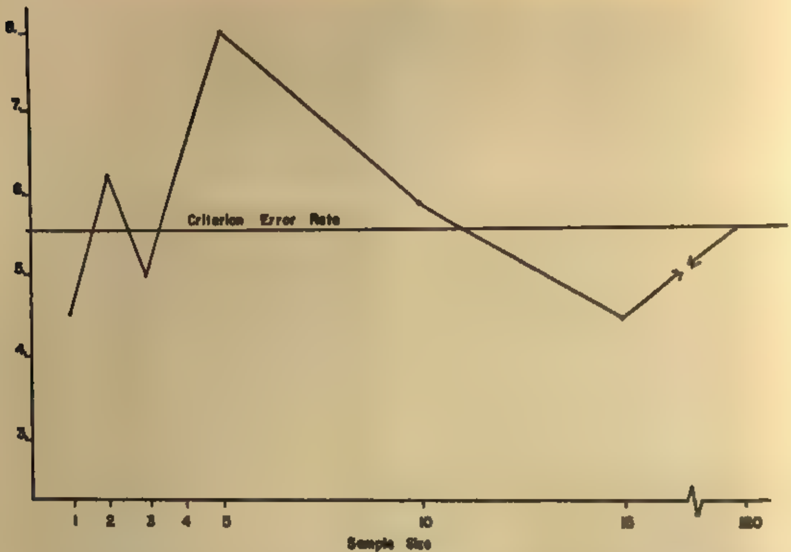


Figure 2. Mean overall error rate estimates as a function of sample size.

case of a 10 per cent error rate criterion, the efficiency of the sample size increases up to size 10. With a 15 per cent or 20 per cent error rate, the efficiency of the sample in terms of the rejection of un-

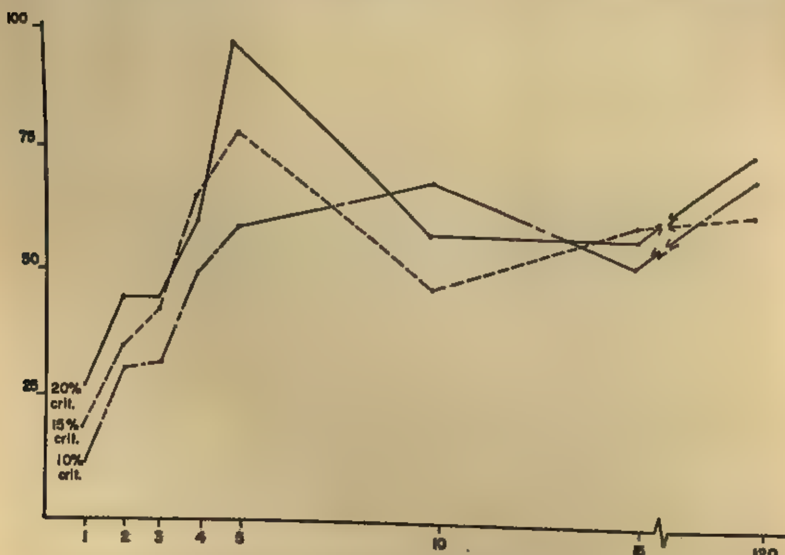


Figure 3. Percentage of the unacceptable items rejected as a function of sample size and rejection criterion.

acceptable items at first increases up to sample size 5 even more rapidly than the curve for the 10 per cent criterion. However, these curves decrease as sample size is increased beyond $N = 5$ and then increased as sample size goes from 15 to 120. Table 3 presents both

TABLE 3

Percentage and Number of Unacceptable Frames that Would be Rejected as a Function of the Size of Sample and the Criterion Error Rate Used for Rejection

Sample		Criterion error rate used for rejection					
		20%[b]		15%[c]		10%[d]	
Size	No.	%	No.	%	No.	%	No.
100	0[a]	100	11[a]	100	24[a]	100	63[a]
1	1	45	5	29	7	19	12
1	2	18	2	17	4	10	6
1	3	18	2	8	2	5	3
2	4	45	5	46	11	38	24
2	5	55	6	38	9	30	19
2	6	36	4	21	5	22	14
3	7	36	4	29	7	29	18
3	8	64	7	42	10	32	20
3	9	36	4	54	13	35	22
4	10	45	5	50	12	25	16
4	11	64	7	79	19	62	39
4	12	73	8	67	16	63	40
5	13	100	11	92	22	68	43
5	14	91	10	58	14	57	36
5	15	100	11	88	21	52	33
10	16	74	7	42	10	56	35
10	17	45	5	38	9	78	49
10	18	64	7	62	15	70	44
15	19	45	5	29	7	51	32
15	20	82	9	58	14	51	32
15	21	45	5	92	22	52	33
120	22 (Summation Sample)	73	8	62	15	68	43

[a] Criterion sample.
[b] 11 frames were unacceptable at this criterion level.
[c] 24 frames were unacceptable at this criterion level.
[d] 63 frames were unacceptable at this criterion level.

the percentage and number of unacceptable items correctly identified by each of the samples of subjects. It should be noted that for a 10 per cent criterion, for example, all percentages are above 50 only

for samples of $N = 5$ or more. However, with a 15 per cent criterion not all the samples of size 15 resulted in 50 per cent or greater correct identifications of unacceptable frames.

Table 4 shows the number and percentage of the acceptable items

TABLE 4

Percentage and Number of Acceptable Frames that Would be Erroneously Rejected as a Function of the Size of Sample and the Criterion Error Rate Used for Rejection

Sample		Criterion error rate used for rejection					
		20%[b]		15%[c]		10%[d]	
Size	No.	%	No.	%	No.	%	No.
100	0[a]	0	0	0	0	0	0
1	1	5.7	17	5.3	15	4.1	10
1	2	2.0	6	1.4	4	0.8	2
1	3	3.4	10	3.5	10	3.7	9
2	4	14.1	42	12.7	36	9.4	23
2	5	12.1	36	11.6	33	9.4	23
2	6	5.7	17	5.3	15	2.9	7
3	7	12.5	37	12.0	34	9.4	23
3	8	10.1	30	9.5	27	6.9	17
3	9	15.8	47	13.4	38	11.8	29
4	10	5.1	15	2.8	8	1.6	4
4	11	28.3	84	25.7	73	21.6	53
4	12	27.6	82	26.1	74	20.4	50
5	13	40.4	120	38.4	109	35.9	88
5	14	26.3	78	26.1	74	21.2	52
5	15	18.9	56	16.2	46	13.9	34
10	16	6.7	20	6.0	17	19.6	48
10	17	17.2	51	16.5	47	43.7	107
10	18	10.8	32	8.5	24	30.2	74
15	19	2.4	7	1.8	5	2.4	6
15	20	3.7	11	2.1	6	6.1	15
15	21	2.7	8	1.8	5	11.0	27
120 (Summation Sample)	22	1.0	3	3.2	9	4.1	10

[a] Criterion sample.
[b] 297 frames were acceptable at this criterion level.
[c] 284 frames were acceptable at this criterion level.
[d] 245 frames were acceptable at this criterion level.

that would be erroneously rejected by each of the samples using the various criteria. These data have been combined with those of Table 3 to produce Figure 4. Figure 4 depicts erroneous rejections

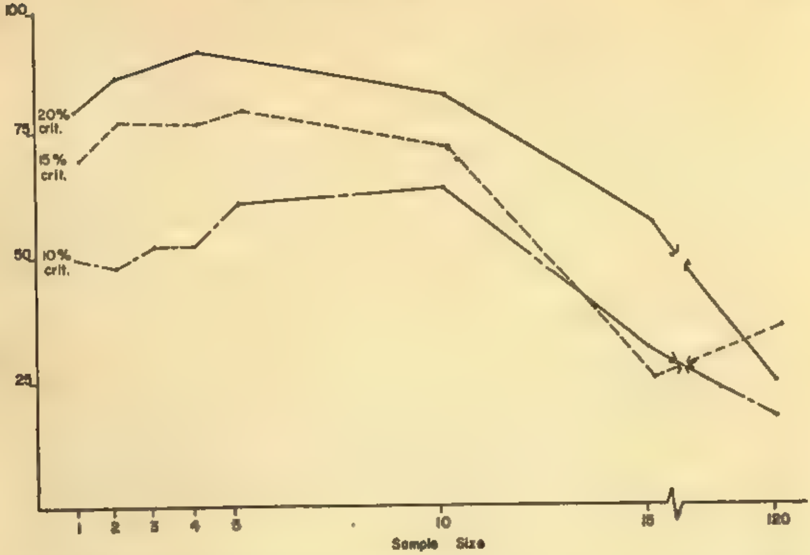


Figure 4. Per cent of rejections made erroneously as a function of sample size and rejection criterion.

in terms of their percentage of all rejections made and thus serves to indicate the "cost" of each correct rejection in terms of incorrectly rejected items.

Discussion

Skewness Effects

The extreme positive skewness of the error rate distribution appears to affect the efficiency of a given sample size in leading to the rejection of unacceptable frames. It seems to have lowered the efficiency of N when the rejection criterion was reduced from a 20 per cent error rate to a 10 per cent error rate. A shift in the rejection criterion of this sort moves the point of rejection to a place in the frequency distribution where considerably more frames are accumulated, since the frames, in general, tend to produce small numbers of errors.

Since the power of the test is lowest for frames which have true error rates that are almost acceptable, the overall efficiency of the tests of frame acceptability is lowered in this situation. One factor, however, works to counteract this reduction in efficiency to some

extent when N is small. It is clear from the formula for the standard error of a proportion $\sigma_p = \sqrt{p(1-p)/N}$ that as the proportion (p) is shifted away from .5, the standard error is reduced. Thus, the standard error of the error rates (proportion of errors) which are observed for borderline unacceptable items will be reduced when the point of rejection is shifted further away from the 50 per cent error rate value. As can be seen from the formula, this effect becomes less important as N is increased.

Overall Error Rates

The mean overall error rates observed for the various sample sizes as depicted in Figure 2 must be considered when interpreting the efficiency curves in Figure 3. Of major concern here is the fact that the mean overall error rate of the samples with $N = 5$ was considerably above that of the criterion. Thus, the efficiency curves in Figure 3 are higher for $N = 5$ than would normally be expected. The same curves are somewhat depressed at the point where $N = 15$ due to the fact that the mean overall error rate happened to be lowest for samples of this size.

Detection of Unacceptable Frames

Two major relationships are illustrated in Figure 3. The first is that in detecting unacceptable frames, samples with $N = 5$ or $N = 10$ can approach quite closely the efficiency of much larger samples. Inspection of Table 3, however, shows that while the *mean* efficiency of the smaller samples is high, variability in efficiency is also quite high. Thus, the programer may be able to approach the efficiency of a very large sample with only a small one, but he runs a definite risk of drawing a very poor sample. It should be noted that even the summation sample leaves a great deal to be desired in the identification of faulty frames.

While it is possible to identify three-fourths or more of the unacceptable frames with samples as small as 10 when a 10 per cent criterion is used, sampling fluctuations are great enough that, in this study, no sample of size 15 was this efficient. The data indicate that when a higher error rate is used as a criterion, smaller samples may identify three-fourths, or more, of the unacceptable frames, and that this could occur with samples as small as 4. While it may

not occur with even larger samples, the chances are greater that it will.

Sample Size and Rejection Criteria

A second relationship, shown in Figure 3, involves sample size and the rejection criterion. The curve for the 10 per cent criterion reaches its first maximum when $N = 10$, while the curve for the 20 per cent criterion reaches its first maximum at $N = 5$. This is due to the fact that with these Ns and these criteria, N is as large as possible at the points where one subject missing a frame will cause it to be rejected. In other words, the lowest possible error rate other than 0 per cent which can be observed in these samples is equal to the rejection criterion at these points. This is not true for samples with slightly smaller or slightly larger Ns. For example, when the criterion for rejection was a 20 per cent error rate and $N = 3$, the only possible observed error rates were 0%, 33%, 67% and 100%. Thus, if a frame was to be rejected, 33 per cent or more of the students had to miss it. This means, in effect, that actions taken in making a decision to accept or reject a frame were the same as if the rejection criterion was shifted to an observed error rate of 33 per cent, while the null hypothesis being tested continued to be that the frame has an error rate less than 20 per cent. The effect of this *de facto* rejection criterion shift is a reduction in the power of the test of this hypothesis. The probability of erroneous rejections (Type I errors) would, however, be reduced. When $N = 5$ in this case, possible observed error rates were 0%, 20%, 40%, 60%, 80% and 100%. Thus, no *de facto* shift in the rejection criterion occurred, since an observed error rate of 20 per cent led to rejection of the item. Increasing N from $N = 5$ to $N = 6$ would, according to the same principle, reduce the power of the test. With $N = 6$, possible observed error rates are 0%, 17%, 33%, 50%, 67%, 83% and 100%. Again, a *de facto* shift of the rejection criterion to 33 per cent would occur and the power of the test would be reduced.

According to this analysis, it would be expected that if more sample sizes had been included in the present study, then each of the curves in Figure 3 would regularly rise and fall as N increased from 0 to 120. Each successive maximum would be a little higher than the last due to increased power brought about by increasing N. Each successive minimum would also be higher due to closer ap-

proximation of the *de facto* rejection criterion to the desired rejection criterion. The maxima would always be at points where the product of the criterion per cent and N is a whole number. For some criteria the maxima will occur with N equal to an integer. For others, N will not always be an integer but will sometimes be an integer plus a fraction. In each of these cases, the observed maximum would be the integer with the fraction dropped. This would be the case for a criterion of 15 per cent. The first maximum for this criterion would theoretically be with $N = 6.7$. The observable maximum would be at $N = 6$, however, since a sample with $N = 6.7$ cannot be obtained.

Efficiency in Terms of Type I Errors

Thus far, the discussion has concerned the efficiency of various sample sizes in terms of the percentage of unacceptable frames rejected. Efficiency should also be evaluated in terms of the extent of Type I errors, the rejection of acceptable frames. Examination of Table 4 reveals that large numbers of acceptable frames were rejected by tests based on the smaller samples. When these numbers are presented as per cents of all rejections made, as in Figure 4, the lower efficiency of smaller sample sizes becomes quite clear. For samples with N less than 15, the best samples led to the rejection of at least one acceptable frame along with every unacceptable frame rejected. The worst rejected as many as nine acceptable frames for every unacceptable frame rejected. The curves in Figure 4 as in Figure 3 have probably been influenced by the mean overall error rates observed with the various sample sizes. Thus, they are somewhat higher where $N = 5$ and somewhat lower at $N = 15$ than would normally be expected. Figure 4 definitely shows, however, that one advantage of using a large pretesting sample is that erroneous rejections are considerably reduced.

It is apparent that the percentage of the rejections made erroneously is also a function of the rejection criterion. For smaller samples, at least, the percentage of rejections made erroneously is lowered when the rejection criterion is reduced from a 20 per cent to a 10 per cent error rate. Such a shift in the rejection criterion results in changing at least three things which would affect the percentage of rejections made erroneously. First, by lowering the rejection criterion, the proportion of acceptable frames in the pro-

gram is reduced and the proportion of unacceptable frames increased. This reduces the probability that an error made by a student will be made on an acceptable frame. Thus, fewer acceptable frames are rejected. Second, the lowering of the criterion places more frames in the "almost unacceptable" category due to the skewness of the error rate distribution. This would tend to increase erroneous rejections. The power of the test in making proper rejections would also be reduced, since more "almost unacceptable" frames would be close to the criterion point. The effect of these changes would be to increase the per cent of rejections made erroneously. A third result of a downward criterion shift would be to reduce the mean standard error of the acceptable frames for reasons already mentioned. Such a reduction in variability lowers the percentage of erroneous rejections. With a large N, however, this effect is considerably smaller. The net effect of all these changes depends considerably on the distribution of error rates in the program.

Hazards of Small Samples

The wide variation in efficiency among samples of a given size, in terms of rejecting unacceptable frames and failing to reject acceptable ones, also obtains where the prediction of overall program error rate is concerned. This is readily seen when inspecting the overall error rate estimates presented in Table 2.

The failure to obtain consistent results with smaller samples in the present study points up a major objection to the use of small pretesting samples. One cannot be confident that a small sample of students will produce individual and overall error rates consistent with those which would obtain in the population for which the program was intended. This objection, along with the fact that erroneous rejections are quite frequent when small samples are employed, must be seriously considered by the planner of a pretesting program. The cost of failing to reject unacceptable frames, of rejecting acceptable frames, and of inaccurately estimating the overall error rate for a program must be balanced against the cost of pretesting the program. When these things are considered the N of the pretesting sample should be set as large as is practical. It should be chosen so that the product of the desired rejection cri-

terion and N is an integer. This will maximize the power of the test.

Sequential Testing

In cases where the program can be pretested with a very large number of students, the programer may wish to adopt a sequential testing procedure. He could use a sample of moderate size and a very strict rejection criterion for the first test. According to our findings, this should identify most of the unacceptable frames although several acceptable frames would be erroneously rejected. The programer would then restrict his frame analyses based on the remainder of the pretesting data to those frames rejected by the first test. Here the normal rejection criterion would be employed with a large N. This second test would reduce the probability of erroneous rejections and simultaneously eliminate the necessity of carrying out large-scale item analyses for the majority of the program frames since most of these frames will have error rates considerably below the rejection criterion.

Summary

In spite of the common practice, in developing programed learning materials, of using small samples of students from the target population to accept or reject frames, there has been no examination of the implications of this practice. This study relates the problem to the problem of the statistician who is testing a large number of hypotheses. The concepts of rejection level, Type I and II errors, and the statistical concept of the power of a test are applied. The empirical nature of the study is important since it is characteristic of the errors made to be intercorrelated and to form a skewed distribution with a mean that departs substantially from .5. Twenty-one independent samples of seven different sizes and three per size were drawn from student worksheets used in learning from an algebra program based upon the UICSM curriculum. The hazards of small samples (up to $N = 15$) with rejection criterion levels of 10%, 15% and 20% were examined. Wide variations in efficiency among samples of a given size were observed both in terms of (a) rejection of acceptable frames, and (b) failure to reject unacceptable ones. Coupled with the inconsistency of small pretesting samples is the high frequency of erroneous re-

jections. It was recommended that pretest samples be both as large as practical and chosen, so that the product of the desired rejection criterion and N are integers, so as to maximize the power of the test. In some cases a sequential sampling procedure may be advisable.

REFERENCES

Beberman, M. and Stolurow, L. M. Comparative Studies of Principles for Programing Mathematics for Programed Instruction. Semi-annual Report for Description of the Modes, Schools and Classes Involved in the 1962–63 Tryout for UICSM Programs. Urbana, Illinois: University of Illinois, 1963.

Holland, James G. Research on Programing Variables. *In Teaching Machines and Programed Learning, II: Data and Directions.* R. Glaser (Ed.) Washington, D. C.: National Education Association, 1965. Pp. 66–117.

Stolurow, L. M. Idiographic Programming. *NSPI Journal* (National Society for Programed Instruction), 1965, 4, 10–12.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 661-664.

THE REGRESSION TOWARD THE MEAN IN MMPI, CALIFORNIA PSYCHOLOGICAL INVENTORY AND SYMPTOM CHECK LIST

R. M. JURJEVICH
NP Clinic, 3415 Dispensary, Lowry AF Base, Denver, Colorado

THE problem of regression toward the mean of retest scores appears insufficiently explored. It has been demonstrated with different instruments and populations. On the other hand some test scales and some populations do not show the regression effect (*cf.* Dahlstrom and Welsh, 1960, p. 384). Some investigators also seem to take it for granted that the regression toward the mean would be more marked with high initial scores than the low ones. The supposition is logical as there is more latitude for changes in extreme scores than in the moderately elevated ones. However, the statistical verification does not bear out that supposition, as a general rule, for the 45 scales investigated.

Method

In studies preliminary to evaluation of individual and group psychotherapy programs (Jurjevich, 1966a, 1966b), the AF male personnel were retested after a short interval (SI) and long interval (LI). The mean interval for the SI group was about 10-12 days, and for the LI group about 180–210 days. The SI group had 51 *Ss* with MMPI retests, 43 with California Psychological Inventory (CPI), and 52 for Symptom Check List (SCL). The corresponding number in the LI group were: 55, 50, 56. The range of initial scores for each scale in the SI and LI groups was divided at the midpoint into high scores (HS) in the higher two quartals and low scores (LS) in the lower two. The differences between the test and retest scores in HS and LS were calculated on the *t* tests.

Results and Discussion

Table 1 shows that among the 45 scales employed, significant differences in amounts of regression, *d*, were found on seven scales in the SI group, 25 in the LI, and 23 in the combined SI + LI group. When the criterion of significance is widened to include scales significantly different at the level of $p = .06 - .10$, the corresponding number of significant differences becomes 10, 28, and 25 in the SI, LI, and SI + LI groups respectively. Apparently the regression toward the mean is more pronounced on some scales while being insignificant on others. The regression towards the mean is also more marked after a longer interval between test and retest than after a short interval, although this does not hold true in all cases. There are 10 scales (MMPI: *Pd, Mf, At, Hv;* CPI: *Do, Re, Sc, Cm, Py;* SCL raw[1]) which show no differences in regression toward the mean in any of the three conditions (SI, LI, and SI+LI). Some scales exhibit significantly different regression effects in all three conditions (MMPI: *F, Pa;* CPI: *Sp, Fx*). Two scales reach significant differences, on the SI only (MMPI: *Cn, Jh*), six on the LI only (MMPI: *L, K, Ma, Hc;* CPI: *To, Gi*), and four in the SI + LI only (MMPI: *Ho, Pd* + .4K; CPI: *Wb*; SCL weighted). Sixteen scales show differential regression effects on LI and SI + LI, but not SI (MMPI: *Hs, D, Hy, Pt, Sc, Si, R, Hs + .5K, Pt + K, Sc + K, Ma + .2K;* CPI: *Ie, Cs, Sa, Ac, Ai*). Two of the scales (CPI: *So, Fe*) have significant differences in the SI and SI + LI, but not the LI, and one scale (CPI: *Sy*) in the LI and SI, but not the SI + LI. It is evident that the regression effects depend largely on the individual characteristics of a scale, so that no general rule as to regression effects can be postulated for the self-assessment scales employed in this study.

The number of *S*s falling into HS and LS categories is relatively even on some scales and uneven on others. In some cases these uneven *N*s may act as suppressors of the significant regression effects, as the calculations for *t* were done by Walker and Lev (1953) formula 7.23. If it be considered that a difference of 51 per cent or more between *N*s of the HS and LS group might be large enough to influence the level of significance on the *t* test, it is found that there

[1] The "SCL raw" represents the number of symptoms admitted, the "SCL weighted" indicates the degree of trouble experienced.

TABLE 1

The Differences of Means (d) of the High (HS) and Low Scores (LS) on Initial Tests in the Short (SI) and Long (LI) Test-retest Intervals

Scale	*d* of HS − LS SI	LI	SI + LI
MMPI raw			
L Lie	.73	1.17*	.95*
F Validity	3.34*	6.29**	4.62**
K Correction	−.30	1.90**	.45
Hs Hypochondriasis	.42	5.76**	3.48**
D Depression	−1.32	4.99*	2.04[a]
Hy Hysteria	.51	4.06**	2.35**
Pd Psychopathic deviate	−1.47	1.59	.11
Mf Masculinity femininity	.30	.95	.58
Pa Paranoia	1.87[a]	4.80**	3.26**
Pt Psychasthenia	1.93	7.02**	4.56**
Sc Schizophrenia	.13	15.58**	6.96**
Ma Hypomania	.47	2.21*	1.37
Si Introversion	.00	4.62[a]	3.07*
At Anxiety, Taylor	−1.34	1.44	−.20
Cn Control	3.30*	−.14	1.49
Hc Hostility control	−.17	3.28*	1.59
Ho Hostility, Cook	−2.07**	4.27*	1.16
Hv Overt hostility	.66	.86	.77
Jh Judged hostility	−2.15*	.41	−.75
R Repression	1.13	2.94*	2.01**
MMPI *K* corrected			
Hs + *.5K*	−1.44	6.09[a]	7.41**
Pd + *.4K*	2.26	5.65	4.73*
Pt + *K*	1.12	15.75**	7.10**
Sc + *K*	3.39	18.34**	12.16**
Ma + *.2K*	1.73	6.11*	4.07*

Scale	*d* of HS − LS SI	LI	SI + LI
CPI raw			
Do Dominance	.38	2.54	1.03
Cs Capacity for status	.55	3.90**	2.21*
Sy Sociability	2.67*	3.89*	3.31
Sp Social presence	2.69[a]	6.71**	4.92**
Sa Self-acceptance	1.99	2.70*	2.31**
Wb Well being	.43	3.43	4.01**
Re Responsibility	−.58	1.59	.68
So Socialization	−3.51**	−.06	1.87[a]
Sc Self-control	−.66	2.36	1.12
To Tolerance	.75	4.08*	2.39
Gi Good impression	.76	3.65[a]	1.74
Cm Communality	.67	2.64	1.01
Ac Achievement via conformance	2.25	6.34**	4.37**
Ai Achievement via independence	.46	3.72**	2.24**
Ie Intellectual efficiency	.30	5.97*	3.49*
Py Psychological mindedness	1.17	3.63	2.53
Fx Flexibility	2.47**	2.89*	2.91**
Fe Femininity	1.24[a]	1.56	1.35*
Symptom CHECK LIST			
raw	.74	3.13	2.34
weighted	1.60	19.24	9.43*

* t = .05 significant difference.
** t = .01 significant difference.
[a] t = .06–.10 significant difference.

are 39 scales having disparities in *N*s that large. Twenty three of these scales show a significant difference between means of the HS and LS groups in spite of the suppressing effect of uneven *N*s. The remaining sixteen scales represent less than 12 per cent of the total number investigated, indicating that disparities in *N*s did not contribute substantially to the results obtained.

In conclusion it may be pointed out that if the scales with significant regression toward the mean were used in assessing personality changes, it would be essential to equate the groups on initial scores to avoid the contaminating effects of regression. Such a procedure does not appear important for scales not showing different regression effects of the HS and LS.

Summary

Groups of about 50 AF males were retested after intervals of about 10 and 180 days. These groups were divided into low and high scorers and differences in regression toward the mean determined. More than half of 45 scales show no different regression with high and low initial scores. Other scales show individualized patterns, depending partly on the length of interval between test and retest.

REFERENCES

Dahlstrom, C. W. and Welsh, G. S., *An MMPI Handbook*, Minneapolis: University of Minnesota Press, 1960.

Jurjevich, R. M., Short Interval Test-retest Stability of MMPI, California Psychological Inventory, Cornell Index and a Symptom Check List. *Journal of General Psychology*, 1966, 74, 201–206. (a)

Jurjevich, R. M., Non-specific Therapy or Spontaneous Remission. 1966, to be published. (b)

Walker, Helen, and Lev, J. *Statistical Inference*, New York: Holt, 1953.

Educational and Psychological Measurement
1966, 26, 665-674.

THREE EQUIVALENT FORMS OF A SEMANTIC DIFFERENTIAL INVENTORY[1]

LOLAFAYE COYNE and PHILIP S. HOLZMAN
The Menninger Foundation

In the course of studying people's reactions to hearing their own voices, we constructed three equivalent forms of a semantic differential inventory. We required that the forms be short enough to be filled out in less than one minute, yet long enough to yield stable scores. We used the forms to trace the momentary attitude changes towards one's own voice at three successive but closely spaced times. Our experimental purpose, that of obtaining independent yet valid assessments of momentary attitude towards one's own voice, directed that the items on the three forms be equivalent but not identical to each other. The instruments finally constructed proved to be most satisfactory and we present the forms here for the use of any colleagues who might wish to employ them in their own work.

The forms, of course, were constructed to yield information about the subject's own voice, and the concept to be rated by the polar adjectives is "my voice." Nevertheless, there may be other concepts for which the items could be appropriate. Some students may wish to test whether these forms maintain their equivalence when they are used to describe other concepts. We are also suggesting our method of inventory construction as one that can assure equivalence. Too often investigators cull items from the Osgood, Suci, and Tannenbaum book (1957) without pretesting their relevance or their factor structure for the concept they wish to study. We therefore believe that whether or not the forms are appropriate for as-

[1] This study is part of a larger study on self-confrontation and is supported in part by Public Health Service Research Grant MH 07962, National Institute of Mental Health.

sessing concepts other than "my voice," the method of test construction recommends this presentation.

Method

From the Osgood, Suci, and Tannenbaum book (1957) we first chose a pool of 70 items applicable on an *a priori* basis to the concept "my voice." We arranged these 70 items in four orders: (1) a first random order, (2) reverse of the first random order, (3) a second random order, and (4) reverse of the second random order. Twenty-nine subjects were assigned randomly to each order making a total sample of 116 subjects. The subjects were sophomore, junior, and senior students at a midwestern university.[2] They ranged in age from 19 to 24. Means, standard deviations, and the intercorrelation matrix for all 70 items over all 116 subjects were computed.

We obtained a principal axis factor analysis of the intercorrelation matrix. The lower bounds of the multiple correlation coefficients were used as diagonal values. Four factors were extracted: the first contained 47 per cent of the common variance; the second, 30 per cent; the third, 12 per cent; and the fourth, 11 per cent. These factors accounted for 17 per cent, 11 per cent, four per cent, and four per cent respectively of the total variance. There was too sharp a drop in the percentage of variance accounted for by succeeding factors to warrant further extraction of factors.

The four factor solution was rotated by the normal varimax rotation method, and after rotation, Factor I accounted for 33 per-cent of the common variance; Factor II, 30 per cent; Factor III, 24 per cent; and Factor IV, 14 per cent of the common variance. Of the total variance, Factor I accounted for 12 per cent; Factor II, 11 per cent ; Factor III, nine per cent; and Factor IV, five per cent.

We then chose groups of items that loaded almost exclusively on each of the four factors. We were able to choose 28 items for Factor I, 19 items for Factor II, 14 items for Factor III, and nine items for Factor IV. The items and their loadings are given in Table 1. In Table 1 defining loadings are considered to be loadings of .20 or above, but in our own actual procedure of selecting items we considered .30 or above to be a minimum loading.

---

[2] We gratefully acknowledge the assistance of Mrs. Winifred Siegel who administered the forms and did all of the preliminary scoring.

TABLE 1

Normal Varimax Factors

Item No.	Item	I	II	III	IV	h^2
46	active–passive	65	18	05	34	57
2	sociable–unsociable	64	−15	11	−17	47
11	positive–negative	62	17	29	07	50
41	sharp–dull	61	09	−03	22	43
17	bright–dark	61	−33	13	05	50
25	constrained–free	−61	−02	−05	15	40
9	successful–unsuccessful	56	03	45	06	53
14	fresh–stale	56	−03	34	18	46
42	blatant–muted	56	20	−07	16	38
1	optimistic–pessimistic	55	−07	09	−13	33
6	graceful–awkward	54	−08	46	05	51
20	clear–hazy	53	−04	29	29	46
34	wide–narrow	51	34	23	15	46
56	refreshed–weary	49	−14	38	26	47
57	colorful–colorless	48	14	48	20	52
62	near–far	48	−19	14	03	28
38	constricted–spacious	−47	02	−05	20	27
43	fast–slow	47	−15	−06	18	28
58	interesting–boring	46	25	39	17	45
13	believing–skeptical	45	−17	16	−13	27
68	public–private	43	13	−06	21	24
40	hot–cold	37	09	17	−05	18
55	savory–tasteless	35	01	28	09	21
35	long–short	35	11	10	14	16
63	tangible–intangible	30	06	26	−06	16
27	opaque–transparent	−29	22	08	−17	17
69	humble–proud	−26	−24	−04	−12	14
48	red–green	24	−05	−07	18	10
37	rugged–delicate	05	82	03	−04	68
28	masculine–feminine	−04	79	−03	−03	63
31	heavy–light	00	79	03	−01	63
32	hard–soft	−04	72	−22	07	58
33	thick–thin	−03	72	23	04	57
30	deep–shallow	18	66	21	22	56
23	strong–weak	39	60	30	13	61
54	youthful–mature	−03	−59	−02	19	38
36	large–small	41	58	−05	09	52
44	sweet–bitter	13	−54	46	03	52
5	light–dark	25	−52	06	05	34
21	sacred–profane	02	−49	14	−06	27
29	rough–smooth	−22	48	−43	−11	48
24	severe–lenient	−01	43	12	28	28
64	wet–dry	20	30	27	25	26
26	serious–humorous	−09	28	14	−02	11
3	kind–cruel	16	−27	27	−02	17
70	objective–subjective	09	26	01	18	11
67	sophisticated–naive	20	23	19	13	14
19	nice–awful	25	−07	67	03	52
7	pleasurable–painful	38	06	63	09	56
8	beautiful–ugly	19	−03	58	19	41
22	good–bad	34	13	53	−08	43

TABLE 1 Continued

Item No.	Item	I	II	III	IV	h^2
16	calm–agitated	−14	06	53	−07	31
15	rich–poor	25	24	52	13	41
18	sweet–sour	12	−39	49	−04	41
10	important–unimportant	28	15	47	03	32
4	clean–dirty	05	−36	44	14	34
66	formed–formless	24	31	44	17	37
12	reputable–disreputable	30	−21	38	−16	31
50	cautious–rash	−09	−17	34	04	16
47	angular–rounded	03	−04	−32	08	11
49	stable–changeable	−10	19	31	17	17
53	unusual–usual	−01	−04	−13	74	57
39	complex–simple	−06	07	13	69	50
52	new–old	19	03	12	59	40
61	ornate–plain	15	20	21	45	31
51	straight–curved	09	29	11	−40	26
65	competitive–cooperative	23	31	−15	39	33
45	varied–repetitive	17	03	35	38	30
59	pungent–bland	26	21	11	35	25
60	sensitive–insensitive	−17	−26	24	34	27

Inspection of the items after the rotation solution suggests that Factor I be called an *activity factor*, Factor II a *potency* factor, Factor III an *evaluative factor*, and Factor IV a *complexity* factor. Factors I, II, and III are consistent with the factor analyses previously reported by Osgood, *et al.* (1957), although the evaluative factor, which in most other studies accounts for the largest portion of the variance, was in our solution the third highest factor in terms of per cent variance.

To construct the three equivalent sets of items for each factor, we originally planned to select individual sets of three items which: (1) are highly intercorrelated, (2) have similar *patterns* of intercorrelations, and (3) have nearly equal means and standard deviations. We then proposed to assign these three items randomly to each form, and continuing this method of item selection, we hoped to obtain five items per factor for each of the three forms. The relatively small pool of items for each factor, however, prevented our using an *individual* item matching procedure. Therefore, we required that the items for each factor be matched in such a way that the *average* of their means, standard deviations, and the factor loadings of each item are approximately equal across the three forms. We also required that items have factor loadings of at least .30. On the basis of these criteria, we were able to select for each

form seven items for Factor I,[3] four items each for Factors II and III. Factor IV contained too few item to warrant inclusion.

Table 2 presents the items selected for the three forms to represent Factor I, the activity factor, their means, standard deviations, and factor loadings, as well as the mean, standard deviation, and factor loading for each form. The maximum difference in means among the three forms was .05; the maximum difference in standard deviations was .02; the maximum difference in factor

TABLE 2

Factor I—Activity

Form	Item No.	Item	Mean	Standard Deviation	Factor Loading
A	2	sociable–unsociable	2.68	1.34	.64
	11	positive–negative	2.95	1.27	.63
	9	successful–unsuccessful	3.00	1.01	.56
	14	fresh–stale	3.07	1.19	.56
	34	wide–narrow	3.53	1.28	.52
	13	believing–skeptical	3.12	1.56	.45
	68	public–private	3.94	1.45	.43
		Average	3.18	1.31	.55
B	41	sharp–dull	3.25	1.12	.62
	−25	free–constrained	2.80	1.58	.61
	42	blatant–muted	3.78	1.13	.56
	20	clear–hazy	2.91	1.38	.53
	62	near–far	3.36	1.29	.48
	−38	spacious–constricted	2.99	1.30	.47
	63	tangible–intangible	3.37	1.43	.30
		Average	3.21	1.33	.52
C	46	active–passive	2.98	1.40	.65
	17	bright–dark	3.26	1.28	.61
	1	optimistic–pessimistic	3.10	1.49	.55
	6	graceful–awkward	3.56	1.30	.54
	56	refreshed–weary	3.03	1.22	.49
	43	fast–slow	3.32	1.45	.47
	58	interesting–boring	2.90	1.17	.46
		Average	3.16	1.33	.54

[3] Out of a pool of 25 items with factor loadings greater than or equal to .30, 21 were chosen for the three forms. Item #57, "colorful-colorless," was omitted from Factor I because it had an equal loading on Factor III. Items #40, "hot-cold;" #55, "savory-tasteless;" #35' "long-short," were also omitted from Factor I. With the exception of item #9, these three items had the lowest standard deviations of any of the items available for Factor I, indicating less discriminative power. Their factor loadings, too, are the lowest loadings among the pool of items with the exception of item #63.

loadings was .03. The average intercorrelation among the items for the first form was .32; for the second form was .25; and for the third form was .33. The average intercorrelation of the items for the first form with the items in the second form was .28; the items in the first form with the items in the third form was .33; and the items in the second form with the items in the third form was .26.

Table 3 presents the items for each of the three forms for Factor II, the potency factor, together with means, standard deviations, and factor loadings for both items and forms. For the three forms on Factor II, the maximum difference among means is .07; among standard deviations is .06; and among factor loadings is .03. The average intercorrelation among items is .41 for the first form; .32 for the second form; and .32 for the third form. The average intercorrelation between items of the first form and the second form is .42; between the first form and the third form is .38; and between the second form and the third form is .34. Of the 14 available items for Factor II having loadings of .30 or above, 12 were used in the three forms.[4]

TABLE 3

Factor II—Potency

Form	Item No.	Item	Mean	Standard Deviation	Factor Loading
A	37	rugged–delicate	3.46	1.32	.82
	32	hard–soft	3.88	1.54	.72
	23	strong–weak	3.00	1.49	.60
	−5	dark–light	3.12	1.32	.30
		Average	3.36	1.42	.64
B	31	heavy–light	3.53	1.49	.79
	30	deep–shallow	3.25	1.38	.66
	−54	mature–youthful	2.91	1.68	.59
	24	severe–lenient	3.98	1.33	.43
		Average	3.42	1.48	.63
C	28	masculine–feminine	3.18	2.16	.79
	36	large–small	3.53	1.27	.58
	−44	bitter–sweet	3.37	1.01	.54
	−21	profane–sacred	3.33	1.16	.49
		Average	3.35	1.47	.61

[4] Items #29, "rough-smooth," and #33, "thick-thin," were discarded because we were least able to match them on means and standard deviations with other items.

Table 4 presents the items selected to represent Factor III, the evaluative factor, means, standard deviations, and factor loadings for both items and forms. The maximum difference in form means was .04; maximum difference in form standard deviations was .08; and the maximum difference in form factor loadings was .04. We omitted two items from Factor III: items #47, "angular-rounded," and #49, "stable-changeable." These items had the two highest means and the two lowest loadings among the pool of possible items, making them difficult to match. The average intercorrelation among these items for each of the three forms are respectively .26, .21, and .25. The average correlation between the items of the first form and the items of the second form is .29; between the first form and the items in the third form .26; and between the items in the second and third forms .27.

TABLE 4

Factor III—Evaluative

Form	Item No.	Item	Mean	Standard Deviation	Factor Loading
A	7	pleasurable–painful	2.73	1.06	.63
	8	beautiful–ugly	3.60	0.96	.58
	4	clean–dirty	2.72	1.25	.44
	50	cautious–rash	3.35	1.41	.34
		Average	3.10	1.18	.51
B	22	good–bad	2.91	1.34	.54
	18	sweet–sour	3.46	1.05	.49
	10	important–unimportant	2.99	1.20	.47
	66	formed–formless	2.95	1.26	.44
		Average	3.08	1.22	.49
C	19	nice–awful	2.96	1.22	.67
	16	calm-agitated	3.38	1.47	.53
	15	rich–poor	3.27	1.12	.52
	12	reputable–disreputable	2.87	1.19	.38
		Average	3.12	1.26	.53

Although the selection of items for the three equivalent forms was done on a form matching, rather than individual item matching, basis it can be seen that for all three factors the form means, form standard deviations, form factor loadings, and form average intercorrelations are strikingly close.

We tested the equivalance of the three semantic differential in-

ventory forms in an experiment employing 20 *Ss* in an experimental group and 20 *Ss* in a control group. *Ss* in the experimental group tape recorded a standard passage and then filled out Form A of the semantic differential (condition I). Then they listened to a five-second sample of their recorded voices and immediately completed Form B of the semantic differential (condition II). Five minutes later, after an interpolated audiometric test, *Ss* completed Form C of the semantic differential (condition III). *Ss* in the control group followed the identical procedure except that they completed Form A after each of the three conditions.

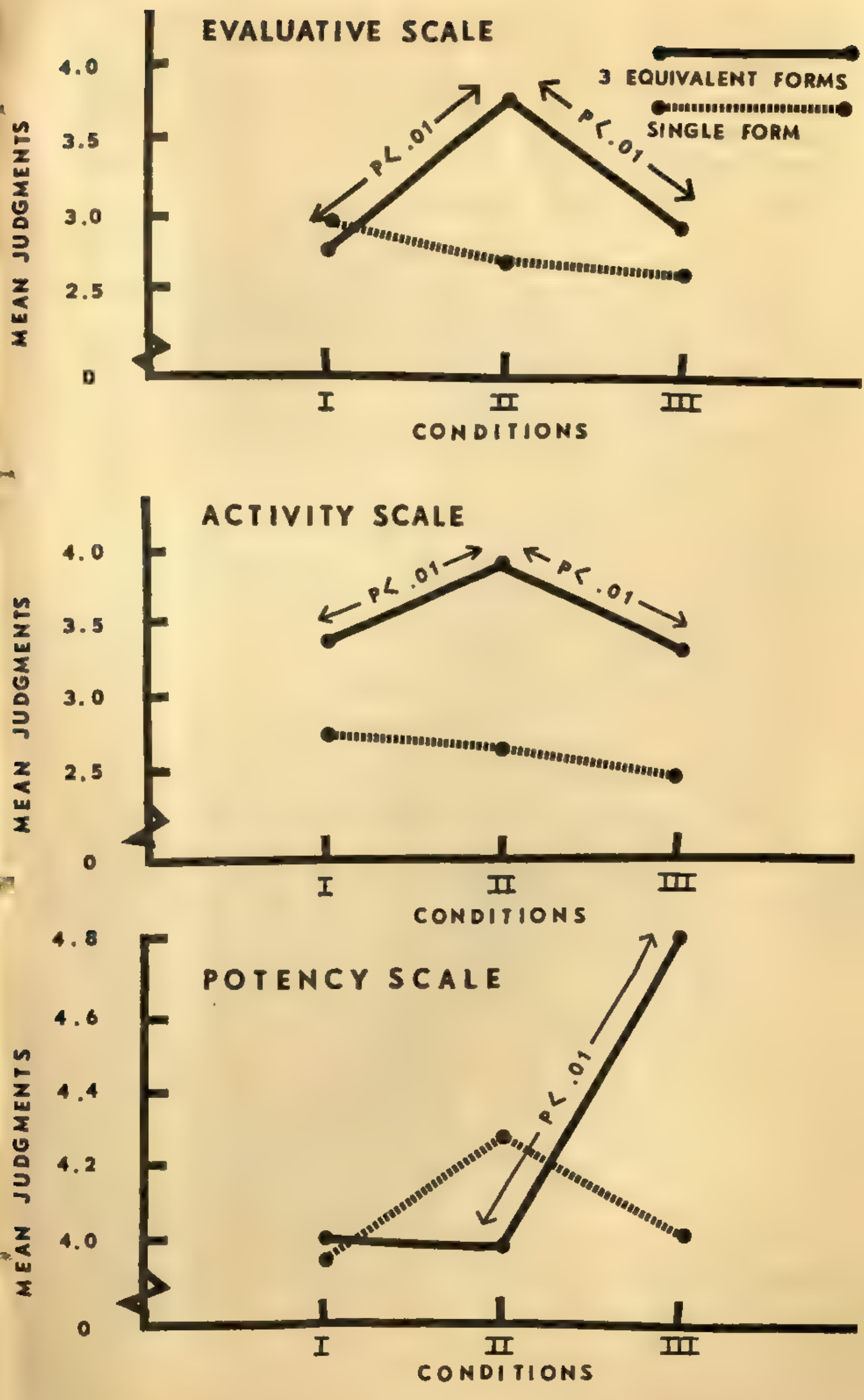
Figure 1 presents the results. *Ss* in the experimental group—those who completed the three equivalent semantic differential forms—show an experimental effect of listening to their own voices: there is a significant shift in the evaluative and activity scales from conditions I to II, and in the potency scale from conditions II to III. These shifts are attributable to the experience of listening to their own voices. For *Ss* who completed the same form for each of the three conditions no such experimental effect appears. The psychological implications of these results are discussed in another paper (Holzman and Rousey, in press). Here we wish only to indicate that repeated responses to the same semantic differential form reflect primarily the repeat reliability and error variance of that form, making the form relatively insensitive to momentary attitudinal changes. The measurement of such fleeting shifts in attitude by means of a semantic differential inventory would therefore require construction of equivalent forms, in the manner suggested here, to control the serial effect of repeatedly using the same form.

Summary

We constructed three equivalent forms of a semantic differential inventory relevant to the concept, "my voice." The effectiveness of the use of equivalent forms for measuring quick attitudinal shifts

Caption for Figure on opposite page.

Figure 1. Mean semantic differential judgments of "My Voice" on the evaluative, activity and potency scales for 3 conditions of judgment (before, immediately after and 5 minutes after listening to their own voices) by 2 groups of *Ss*: those using 3 different but equivalent forms and those using the same form for each condition.



EVALUATIVE SCALE
3 EQUIVALENT FORMS
SINGLE FORM
P < .01
P < .01
MEAN JUDGMENTS
4.0
3.5
3.0
2.5
0
I
II
III
CONDITIONS
ACTIVITY SCALE
P < .01
P < .01
MEAN JUDGMENTS
4.0
3.5
3.0
2.5
0
I
II
III
CONDITIONS
POTENCY SCALE
P < .01
MEAN JUDGMENTS
4.8
4.6
4.4
4.2
4.0
0
I
II
III
CONDITIONS

was demonstrated. Whether the forms maintain their equivalence when used to describe the meanings of other concepts is a matter for empirical test. We recommend this method of semantic differential inventory construction as one that is more precise than simply culling items from a large pool on an *a priori* basis.

REFERENCES

Holzman, P. S. and Rousey, C. The Voice as a Percept. *Journal of Personality and Social Psychology*, in press.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. *Measurement of Meaning*. Urbana, Illinois: University of Illinois Press, 1957.

Educational and Psychological Measurement
1966, 26, 675-690.

THE EFFECT OF SITUATIONAL VARIABLES ON THE MEASUREMENT OF ACHIEVEMENT MOTIVATION

RICHARD E. CARNEY
California Western University[1]

The present paper will consider some of the traditional distinctions between objective and projective measures of achievement motivation, will present evidence which casts doubt on such distinctions, and will offer a tentative basis for integrating those two approaches into a common framework. The discussion will be limited for purposes of economy to a single pair of measures which purport to measure a single motive; however, it is anticipated that some generalization will prove possible.

Our example of an objective measure of achievement motivation is the Achievement-Orientation (*AO*) scale from the California Psychological Inventory (CPI) (Gough, 1957). Our example of a projective measure is the need Achievement measure (*n Ach*) from the Thematic Apperception Test (TAT) (McCelland, Atkinson, Clark, and Lowell, 1953). Each of these measures has substantial evidence for validity (Atkinson, 1958; Carney, 1961, 1963a; 1964; Carney, and McKeachie, 1963), yet the association between them has been elusive as has so often been the case when objective and projective measures of the same variable are compared (Fiske, 1963). Such lack of relationship and other considerations have lead McClelland (1958) to conclude that the self-descriptive and the fantasy measures do not measure the same thing.

A brief bit of preliminary theorizing may help to resolve some confusion in the area of projective versus objective measurement.

---

[1] Formerly at Indiana University (Northwest Campus) and Drake University.

After which consideration will be given to empirical correlations which have been computed between *n Ach* and *AO* under a wide variety of conditions, and the latter portion of this paper will be devoted to a discussion of the variables which affect the *n Ach-AO* correlation (r).

Concepts of Personality

Personality exists only as a process of interaction between the organism and its environment (Fiske, 1963). Every behavior, however it may be observed, has originated in this same process. It is time we stopped even talking as if there were some independently functioning "part" of personality to be projected into behavior. Any concept of a motive capable of indeterminate freedom of action is the true heir of vitalism, and we should be, by now, finished with such concepts. Unfortunately, this is not the case (Carney, 1963b, Immergluck, 1964). It has become almost traditional to assert that the projective situation permits "greater freedom of response" (no real indeterminancy intended) (*e.g.* Lindzey, 1961, pp. 42–46). In practice such statements refer to the fact that a nebulous pattern of stimuli leads to a highly variable set of responses.

A strictly deterministic model requires that *all* behavior upon which inferences of motivation are based be a joint interactive function of the organism and its environment. If we accept Atkinson's (1958) formulation of Motivation $= f$ (Motive $\times$ Expectancy $\times$ Incentive) and his definition of a motive as a relatively stable disposition to act, then the behavior we observe (aroused motivation) is by definition jointly determined by the action of a particular set of stimuli with a particular organismic structure. Those working with the *n Ach* measure have been acutely aware of situational influences which can and do change the responses to the TAT (Atkinson, 1958). Lindzey (1961) provides a concise summary of the complexities which enter into the measurement context, and the following discussion will explore a few of these complexities.

Techniques of Measurement

Questionnaires. The questionnaire technique, if well done, eliminates much of the personal interaction from the testing situation. By using a large number of short, well structured, printed items

which refer to presumably familiar and significant life situations, questionnaires tend to control the effective stimulus matrix and to also provide a simple unambiguous response. Hopefully, the *S*'s attention is involved primarily with the presented items and is thereby largely removed from other situational influences. All that is done is to insure a reasonable knowledge of what it was that participated on the environmental side of the O X E interaction. This control can in no way influence the predispositions or motives. It is interesting that the questionnaire is attacked because the responses of the *S*s are "too subjective" (Watson, 1959). Sometimes this critique comes from those who simultaneously argue that fantasy is the proper datum (McClelland, 1958). When faced with a questionnaire item there is little that an *S* can do (if he participates at all) except engage in cognitive and imaginative processes (fantasy?). The need to construct "lie" scales and the discovery of the influence of the Social Desirability factor (Edwards, 1959) testify that a seemingly barren *X* mark can be preceded by remarkable thought processes fully open to unconscious influences. Thus there is nothing intrinsically different about the O side of the O X E interaction, only the *form* of the final response varies. Some investigators like *X*s and some prefer words, but this preference probably indicates more about the investigators than the methods they employ (Fiske, 1963). There is no logical reason to exclude responses to questionnaire items from the valid measures of motivation.

TAT. Behavior elicited by the TAT pictures is no less a function of the O X E interaction. *The major formal difference between the questionnaire and the TAT is that the test stimuli of the TAT provide fewer effective stimuli.* By intention, the test pictures are rendered ambiguous, with only the minimum structure needed to furnish a response setting. Since, by the present analysis, there *must* be stimuli in the measurement situation which provide the activating environment, the net effect is to force the *S* to find cues in aspects of the testing context other than those directly provided by the TAT. It is not immediately obvious why such stimuli should be expected to lead to "fantasy" responses less biased by social factors than responses made in self-description (McClelland, 1958). The crux of the matter here is: Does one "escape" from the determining predispositions learned from society when composing a story about

others, or does one use essentially the same processes as when composing a story about oneself? The object of the tale has superficially changed, but if it *really* had changed there would be little use in analyzing fantasy to infer the motivation of the storyteller. Our stories about ourselves are usually fantastic enough!

Behavior resulting from projective testing is then, derived in the same manner in principle as that from objective testing. There is no inherent reason why it cannot also be a valid measure of motivational differences—provided that some systematic knowledge is available about the stimuli which actually elicited the behavior. (See Holtzman, Thorpe, Swartz and Herron, 1961, for an approach which seems to provide some such knowledge about inkblots.)

A Synthesis

What the fabled difference between objective and projective measurement resolves itself into, then, is that objective techniques produce a relatively standardized stimulus situation which elicit relatively standardized behavior, while projective techniques often produce non standardized behavior which is elicited by non standardized extra-test situational factors. There is no "freedom of response" which produces "interesting and significant deviant responses" in the projective situation. There is simply less control and considerable ignorance as to which stimuli were effective. Table 1 summarizes the above discussion.

TABLE 1
The Influence of Test Stimuli, Context and Motivation on Responses to Objective and Projective Tests

Source of Variation	Type of test	
	Objective	Projective
Test stimuli	high	low
Context	low	high
Motive	constant	constant

Objective-projective as used here is, of course, a dichotomy applied to a continuum, and the *n Ach* measure was selected as the projective example expressly because its users have gone farther

than most in sorting out the confounding effects of situational variables. The moral for measurement seems none-the-less clear, unless we are specifically interested in the reaction of an *S* to the total test context (as we well may be in clinical practice), we are better off to control that context. If we are interested in exploring reactions to the context, there is no easy way out of the necessity of a rigorous and detailed definition of every significant facet of that context (see Lindzey, 1961, pp. 323–328; Fiske, 1963). Such an exploration is ordinarily expensive and quite time consuming. Self-description is *so* much easier, and in principle the same. Why not use it when we can?

Some Implications

TAT responses are assumed to be largely determined by the immediate situation. It is when the situation extra to the test provides cues for achievement motivation that the TAT and questionnaire measures should coincide. *S*s obtained from the classroom should have a general set which reflects the subject matter content, instructor personality, and classroom procedures and this set should systematically contribute to the obtained TAT responses. Questionnaire responses can also be manipulated by varying the set of the *S* (Carney, 1964), but are assumed to be relatively less affected under ordinary circumstances than the TAT responses.

McKeachie (1961) has shown that *S*s high in achievement motivation prefer classroom situations where they can assume some responsibility for their own behavior. Carney (1961) presents evidence of interactions between personality and subject matter on the attitude of the *S* toward a course.

The remainder of this paper documents some of the factors which influence the correlation between CPI and TAT measures of achievement motivation. The following outcomes were expected:

1. A real but highly variable *r* exists between *AO* and *n Ach*.

2. The *AO–n Ach r* will be affected by salient features of the total situation in which the measures are obtained—such as; the motivation of the instructor, and the "atmosphere" of the classroom.

3. The *n* Ach measure will be most sensitive to the situational variables, and it is this sensitivity which most probably accounts for the lability in the relationship between *n Ach* and *AO*.

Procedure

Sample. The *Ss* were drawn from a Mid-Western college sample which is described in detail elsewhere (McKeachie, 1961; Carney and McKeachie, 1963). Over 800 *Ss* in 70 separate classes (including Psychology, French, and Mathematics courses) at two universities were involved.

Measures. n Ach and *AO* measures were derived by standard techniques (Veroff, *et al.;* 1960, Gough; 1957, Carney, 1961). Measures of *n Ach,* and need affiliation (*n Aff*) were obtained from the instructors by using the French Test of Insight, and Bales scoring for group interaction was done for some classes (McKeachie, 1961).

Statistical Procedures. All relationships were plotted on scatter diagrams and visually inspected for factors which might influence the value of a Pearson product-moment correlation (r), such as non-linearity, extreme skewness, mean differences, and directionality of the relationships in sub-groups.

According to the above inspection, rs were computed and compared. The z' transformation was used when averaging or testing for significant differences between rs (Edwards, 1960). The transformed unweighted rs were used for all comparisons between r and other variables.

The unweighted means of the various sub-groups were used for all comparisons between level of motivation and other variables. Analysis of variance and other tests of significance were carried out according to Edwards (1960).

Results

Correlations between n Ach and AO

Total Sample. Table 2 shows the rs for the total sample of males and females. Both the usual r and the mean absolute values of r obtained by averaging the rs from the sub-samples are given. Although there is a low, significant + *n Ach-AO* r for the males, the females have an insignificant $-r$ and the combined r approximates zero. This result agrees well with past findings of no relationship between objective and projective measures of achievement motivation. However, the consistent and relatively large values of the

mean absolute *n Ach-AO r* indicate that there is more to this situation than is immediately apparent. Tables 3, 4, 5, and 6 illustrate this point in detail.

TABLE 2

Correlations between n Ach and AO for Total Sample (University of Michigan, 1957; Drake University, 1962)

Sex	*r*	*df*	Mean absolute *r*	*df*
Male	.109*	400	.305**	203
Female	−.040	456	.351**	362
Total	.075	856	.330**	645

* $p < .05$.
** $p < .01$, two tailed tests.

TABLE 3

Correlations between n Ach and AO for French Classes (University of Michigan, 1957)

Sex	*r*	*n*	Mean absolute *r*	*n* Classes	*df*	Test between *rs*
Males						
Male Inst.	.233	31	———	pooled		MI-FI is n.s.
Female Inst.	−.073	21	———	pooled		
Females						
Male Inst.	.036	93	.427**	9	66	$\chi^2 = 19.86$ $p < .02$ (between classes)
Female Inst.	−.269*	50	.242	7	29	MI-FI is n.s.
Totals						
Male Inst.	.099	124	.373**	9	94	*z* between
Female Inst.	−.185	71	.178	7	47	absolute mean
Male *Ss*	.115	52	.171	pooled	46	*rs* for both
Female *Ss*	−.054	143	.375**	16	95	MI-FI and M*Ss*
All *Ss*	−.020	195	310**	16	141	F*Ss* = 2.1 $p < .05$

Note.—In several classes there were insufficient *ns* to permit the computation of meaningful *rs*. This was true for the males in all French classes. The loss of *Ss* from small classes will make both the number of classes and of subjects change from table to table.
* $p < .05$.
** $p < .01$, two tailed tests for the mean absolute *rs*.

TABLE 4

Correlations between n Ach and AO for Mathematics Classes (University of Michigan, 1957)

Sex	*r*	*n*	Absolute mean *r*	*df*	*n* Classes	Test between *rs*
Males	.254*	85	.261*	61	8	Classes, n.s.
Females	.291	48	.321	27	7	Classes, n.s.
Total	.264**	133	.279**	88	8	M-F, n.s.

* $p < .05$.
** $p < .01$, two tailed tests for the mean absolute *r*.

TABLE 5

Correlations between n Ach and AO for Psychology Classes (University of Michigan, 1957)

Sex	r	n	Absolute mean r	df	n Classes	Test between rs
Males	−.163	75	.577**	56	6	Classes, $\chi^2 = 40.61, p < .01$
Females	.004	137	.325**	119	6	Classes, $\chi^2 = 16.70, p < .01$
Total	−.050	212	.413**	175	6	M-F, n.s.

** $p < .01$, two tailed tests for the absolute mean r.

TABLE 6

Correlations between n Ach and AO for Psychology Classes (Drake University, 1962)

Sex	r	n	Absolute mean r	df	n Classes	Test between rs
Males	.192*	138	.249**	120	6	Classes, $\chi^2 = 17.33, p < .01$
Females	−.217*	130	.363**	121	3	Classes, $\chi^2 = 11.74, p < .01$
Total	−.014	268	.286**	241	6	M-F, $z = 3.31, p. < .01$

Note.—Only 3 instructors participated. One instructor taught 4 of the 6 classes. Only 3 classes had a sufficient n of Female *S*s to permit computation of r; each of these classes was taught by a separate instructor.

* $p < .05$.

** $p < .01$, two tailed tests for the absolute mean rs.

The Effect of Type of Course and Sex

French Classes. Table 3 presents the results of the French classes at the University of Michigan. Since there were both male and female instructors for these classes, the data have been presented accordingly. The n of male *S*s was too small to permit computation of r for each class, so all of the male *S*s were pooled for male instructors (MI) and female instructors (FI) respectively. The female *S*s respond to the female instructors fairly uniformly yielding a homogeneous set of $-r$s between *n Ach* and *AO*. The male instructors produce both greater variability (and a zero order r) and a higher absolute mean value of r. The male *S*s show a pattern of signs for the *n Ach-AO* rs similar to that of the females, but seem to have a higher $+r$ and a lower $-r$ for the MI and FI categories respectively. Again, the r for all *S*s and all classes is near zero. This fact conceals both the effects of the sex of the instructors and the effect of being in an individual class. The substantial value of the absolute mean value of r is also concealed.

Mathematics Classes. Table 4 shows the results for the Mathematics classes at the University of Michigan. In this case the n

Ach-AO rs for both sexes and over all classes were consistent and positive. Both the total *r* and the total absolute mean *r* are significant and in good agreement with each other.

Psychology Classes. Tables 5 and 6 show the results for the Psychology classes at the University of Michigan and Drake University respectively. There is striking disagreement between the *rs* within each table and between tables. As was the case with the MI French classes, total *rs* indicate zero order relationships while mean absolute *rs* show substantial relationships. The signs of the *rs* reverse for the males and females at the different schools, and the *rs* for both males and females are significant at Drake but neither reach significance at Michigan. This latter result is probably due to the fact that one instructor taught four classes at Drake introducing an element of consistency which was lacking at the University of Michigan where six instructors each had one class (one lecture, and two discussion groups).[2]

To summarize the results of this section, the size and direction of the *r* between *n Ach AO* depends on the sex of the *S*, the sex of the instructor, the course content, and the particular class in which the *S* is enrolled. However, the mean absolute *r* is quite stable across the above variables and indicates a population value for *r* between *n Ach* and *AO* of .330.

Effect of Instructor Motivation on the n Ach-AO r

Effect of Single Motives. Table 7 shows the source of some of the variability in the *n Ach-AO r* within courses—the motivation of the instructor from whose class the *Ss* were taken. For this comparison the *n Ach-AO rs* were ordered with the largest $-r$ at the low end of the scale and the largest $+r$ at the high end of the scale. This ordering was then correlated with the motive scores of the instructors. Both male and female *Ss* tend to have $-rs$ in classes where the instructors are highly motivated for either *n* Ach or *n* Aff. When the instructors are low in motivation, the *n* Ach-

[2] R. J. Roper in an unpublished study compared *AO* and *n Ach* in three Introductory Psychology classes at California Western University during the winter quarter of 1965. These classes were taught by separate instructors and contained 33 male and 44 female *Ss*. The absolute mean *r* over classes was .259 ($p < .05$, 68 *df*). Although the sub-group *rs* were statistically homogeneous (total $r = -.195$, $p < .05$, 75 *df*), one class had $r = .121$ (27 *df*) and the average *r* for the other classes was $-.340$ ($p < .05$, 42 *df*). These data extend the basic findings to still another school and section of the country.

AO r tends to be positive. Intermediate levels of instructor motivation produces zero order *r*s. There is an exception to this rule in the French classes where the female *S*s seemed to respond differentially to the sex of the instructor.

TABLE 7

Correlations between Instructor Motives and N Ach AO Correlation (University of Michigan, 1957)

Sex	Courses	*n Ach*	*n* values of *r*	*n Aff*	*n* values of *r*
Males	All	−.481*	16	−.764**	16
Females	Psych & Math	−.530*	13	−.347*	27
	French	.248	14		
Total		−.503*	29	−.529**	43

Note.—Males in the French classes were combined into 2 groups as shown in Table 2. Females in the French classes had a trend opposite to that for the Mathematics and Psychology classes or *n Ach* of the instructor and this *r* is shown separately (.248).
* $p < .05$.
** $p < .01$.

TABLE 8

The Correlation between AO and n Ach as a Function of Bales Scores of Classroom Interaction (University of Michigan, 1957)

Class	Sex	Bales Score: Teacher-Led (TL)	Student-Led (SL)	Warmth (W)	*n* Classes
French (females only)	Male Inst.	−.750*	.054	.342	9
	Female Inst.	−.518	.384	.009	6
Mathematics	M	.304	−.464	−.759*	7
	F	.543	−.614	−.829*	6
Psychology	M	−.786*	.557	.989**	6
	F	−.443	.500	.543	6

Note.—Rank order correlations.
* $p < .05$.
** $p < .01$.

TABLE 9

Correlations between Motives of Instructors and Mean Class Levels of Motive Scoring (University of Michigan, 1957)

Sex	Comparison: *AO* Classes *n Ach* Inst.	*AO* Classes *n Aff* Inst.	*n Ach* Classes *n Ach* Inst.	*n Ach* Classes *n Aff* Inst.	*n* Classes
Males	.202	.100	−.444*	−.326	28
Females	−.109	.277	.494**	−.186	26
Total	.055	.183	.065	−.253*	28

Note.—The male and female *r*s between *n Ach* of instructor and classes are significantly different ($z = 3.27$, $p < .01$). The total *r* between *n Ach* of classes and *n Aff* of instructors is significant by a one tail test if the *n* of correlations is taken to be 54.
* $p < .05$.
** $p < .01$.

Effect of Paired Motives. An analysis of variance was carried out using the transformed *r*s from each sub-group of males and females (one per sex from each class) as the dependent variables and the motivation of the instructor as the independent variable. Classes were grouped according to the standing of the instructor above or below the medians of *n Ach* and *n Aff*. The motive scores were taken two at a time and High-High, High-Low, Low-High and Low-Low classifications were formed. This analysis should be considered only suggestive since a small number of sub-groups entered each classification, and details of the analysis will not be presented. However, two findings were consistent in every comparison and should be investigated further. The pattern of correlations over motive classifications was different for each type of course content producing a significant Course X Motive groups interaction for each combination of motives. Also the average *r*s were near zero for all motive classifications except the Low-Low one where in each combination of motives the average *r* was positive and at a level closely approximating the estimate for the population *r* given above.

Effect of Classroom Procedures on the AO-n Ach r

Bales Scoring and the AO-n Ach r. Nine ratios were obtained from the interaction patterns in each of the University of Michigan classes; Teachers Task (TT), Teacher Assertion (TA), Teacher/Student Ratio (T/S), Student Task (ST), Student Volunteered Acts (SV), Teacher Positive Socio-Emotional Acts (T+), Student Positive Socio-Emotional Acts (S+), and Student Tension (TE) (McKeachie, 1961). The average scores for the French, Mathematics, and Psychology courses showed quite consistent patterns over the nine ratios. On the basis of these patterns and logical considerations, the TT, TA, and T/S ratios were summed to obtain a Teacher Led Class (TL) score; the ST, SA, and SV ratios were combined to obtain a Student Led Class (SL) score; and the T+, S+, and inverted TE ratios were summed to obtain a Warmth (W) score.

Mathematics classes tended to be strongly teacher-led, had the lowest SL scores and were lowest in W. The French classes were strongly student-led, lowest in TL and highest in W. Psychology classes were moderate in every score, but most closely resembled the French classes.

Table 8 shows the *AO-n Ach rs* as a function of the Bales scores. The pattern for the males and females is similar in each case. In both the French and Psychology classes the classes strongly led by the teacher tended to produce —*rs*, while those Psychology classes high in SL and W tended to have +*rs*. The Mathematics classes presented a quite different pattern with the classes higher in TL having +*rs* and the classes high in SL and W having —*rs*.

The overall pattern seems to be that classes which fit the general pattern for their type of course produce +*AO-n Ach rs*, and classes which do not fit the overall pattern produce —*rs*. The small numbers of classes involved make this conclusion most tentative.

Effect of Situational Variables on Level of Motivation.

Instructor Motivation and Level of S Motive Scoring. Table 9 shows the *rs* between the mean level of scoring in a class and the motivation of the instructor. The *AO* means are not significantly affected by the personality of the instructor, but the *n Ach* means are affected. There is a trend (significant at the .05 point by a one tailed test) for classes to have low *n Ach* means if their instructor is high in *n Aff*. A similar trend is found for the males between the *n Ach* score of the instructor and the class mean *n Ach*.

Other Variables and Level of Motive Scoring. The usual finding of higher scoring for females on *n Ach* was observed for both *Ss* and instructors. In the French course the classes with male instructors had $r = -.684$ ($p < .06$, $7df$) between SL and mean *n Ach*. French classes with female instructors had an $r = .929$ ($p < .01$, $4df$) between SL and mean *n Ach*. These two *rs* were significantly different ($z = 3.53$, $p < .001$).

Over all classes a mean rating of satisfaction with the class and class mean *n Ach* were correlated (males, $r = .631$, $p < .01$, $16\ df$; females, $r = -.233$, $29\ df$; male versus female r, $z = 3.58$, $p <$ .001).

Discussion

In the present sample all *Ss* knew they were participating in a psychological research and that they were doing so with the advice and consent of their instructors. It is reasonable to assume that the attitudes and expectations derived in the classroom should play

a role in determining test behavior. The evidence here is that this was the case for *n Ach* and not for *AO*. Such a finding is consistent with the present analysis, but certainly not conclusive. A carefully designed series of parametric experiments are needed to determine the conditions under which the *AO* and the *n Ach* measures are affected by variables outside of the immediate testing stimuli.

A final question remains as to the interpretation of the changes in the *r* between *AO* and *n Ach*. If it is assumed that the meaning of the *AO* measure is relatively constant, but that its application to a given situation depends on the response of the *S* in that situation, then some measure of the situational response is needed. *n Ach* seems to be a good possibility. At the high extreme of motivation, instructors would tend to impose their own motives on the class and by doing so would lessen the chance of class members to satisfy their achievement motives (McKeachie, 1961). At the low extreme of motivation the instructor would tend to leave a vacuum into which the *S*s could rush if they were so motivated. As would be predicted if *n Ach* reflects the degree to which the *S* is responding to his immediate chances of structuring the environment, *n Ach* and *AO* are positively correlated in classes which have low levels of instructor motivation. The *n Ach-AO r*s tend to zero in classes with intermediate levels of instructor motivation. The level of scoring on *n Ach* is lowered (for males) by high levels of instructor motivation.

The Bales scores tend to confirm the above analysis in the Psychology and the French classes. However, the Mathematics classes produced the most consistent +*AO* — *n Ach r*s, and these were the most strongly teacher led classes in the study. The sex of the instructor also seemed to play a role in the French classes. What the *S* expects the interaction pattern in a particular course to be may be important. This is another testable hypothesis in search of a test.

Litwin and Ciarlo (personal communication, 1963) have found that *n Ach* predicts situations in which the *S* has maximum opportunity to structure the task, while a questionnaire measure of achievement motivation predicts tasks which are highly structured. A profitable line of future investigation should be to combine manipulation of the *S*'s test taking set toward the TAT (fostering the belief that *S* can structure or not structure the context within

which the test is given), with variation in the structurability of a performance criterion.

The variable of sex of the *S* remains a strikingly effective but puzzling one. There is a tendency for females to have mirror image patterns to that for males, but this is not always the case. Females in our culture may define achievement motivation as "conforming to the requirements of the defined task." (Carney, 1961, 1963a, 1964). If the situation is high in cues for achievement (and well structured) females may respond most highly in terms of achievement motivation. Evidence for this proposition in the present data include a heightened level of *n Ach* scoring when the instructor is high in *n Ach*, and a frequent tendency to have relationships opposite to those for the males. However, in the case of the Bales scores the obtained *r*s were similar for both sexes. The mysterious female is still with us.

Results reported in the past may be made less confusing by application of the present analysis. McKeachie, (1961, 1963) reports that predictions made for males held true for *n Ach* in the French and Mathematics classes and not in the Psychology classes. Since there is a fairly consistent $+r$ between *n Ach* and *AO* in the French and Mathematics classes and a zero *r* in the Psychology classes the finding by McKeachie is expected. Isaacson (1963) also reports a failure of prediction for *n Ach* which might be explained from the present point of view.

Conclusions

Projective and objective techniques *can* measure the same motive under the proper conditions. What these conditions are needs a great deal of additional research. It may turn out that projective measures can be excellent indicators of the response of the *S* to situational variables.

A promising approach should be to employ *both* objective and projective techniques and to let the relationship between these measures be diagnostic of the effectiveness and meaning of responses obtained in the total testing situation (Fiske, 1963).

REFERENCES

Atkinson, J. W. (Ed.) *Motives in Fantasy, Action, and Society.* Princeton: Van Nostrand, 1958.

Carney, R. E. *An Analysis of University Student Behaviors with Measures of Ability, Attitude, Performance, and Personality.* Ann Arbor: University Microfilms, 1961, Order #61-6325.

Carney, R. E. Achievement Motivation, Anxiety, and Perceptual Control. *Perceptual and Motor Skills,* 1963, 17, 287–292. (a)

Carney, R. E. Man or Man? *SPSSI Newsletter,* July, 1963 (enclosure). (b)

Carney, R. E. Validation of an Objective Measure of Achievement Motivation. Unpublished manuscript (ditto), Indiana University, 1964.

Carney, R. E. and McKeachie, W. J. Religion, Sex, Social Class, Probability of Success, and Student Personality. *Journal for the Scientific Study of Religion,* 1963, 3, 32–42.

Edwards, A. L. *Experimental Design in Psychological Research.* New York: Rinehart, 1960.

Edwards, A. L. Social Desirability and Personality Test Construction. In Bass, B. M. and Berg, I. A. (Ed.) *Objective Approaches to Personality Assessment.* Princeton: Van Nostrand, 1959.

Fiske, D. W. Problems in Measuring Personality. In Wepman, J. W., and Heine, R. W. (Ed.) *Concepts of Personality.* Chicago: Aldine, 1963.

Gough, H. P. *Manual for the California Psychological Inventory.* Palo Alto: Consulting Psychologists Press, 1957.

Holtzman, W. H., Thorpe, J. S., Swartz, J. D., and Herron, W. E. *Inkblot Perception and Personality.* Austin: University of Texas Press, 1961.

Immergluck, L. Determinism-freedom in Contemporary Psychology. *American Psychologist,* 1964, 19, 270–281.

Isaacson, R. L. Need Achievement, Need Affiliation, Anxiety and Performance. Paper presented at the 1963 American Psychological Association Convention, Philadelphia, Pa.

Lindzey, G. *Projective Techniques and Cross-cultural Research.* New York: Appleton-Century-Crofts, 1961.

Litwin, George H. and Ciarlo, James A. Achievement Motivation and Risk—taking in a Business Setting. Unpublished manuscript. Harvard University, 1964.

McClelland, D. C. Methods of Measuring Human Motivation. In Atkinson, J. W. (Ed.) *Motives in Fantasy, Action, and Society.* Princeton: Van Nostrand, 1958.

McClelland, D. C., Atkinson, J. W., Clark, R. A., and Lowell, E. L. *The Achievement Motive.* New York: Appleton-Century-Crofts, 1953.

McKeachie, W. J. Interactions of Student Personality, Teacher Characteristics, and Performance. Paper presented at the 1963 American Psychological Association Convention, Philadelphia, Pa.

McKeachie, W. J. Motivation, Teaching Methods, and College Learning. In Jones, M. R. (Ed.) *Nebraska Symposium on Motivation,* 1961 Lincoln, Nebraska: University of Nebraska Press, 1961.

Roper, R. J. The Objective and Projective Measurement of Achievement Motivation, A Comparative Study. Unpublished manuscript, California Western University, 1965.

Veroff, J., Atkinson, J. W., Feld, S., and Gurin, G. The Use of Thematic Apperception to Assess Motivation in a Nationwide Interview Study. *Psychological Monographs*, 1960, 74, Whole #499.

Watson, R. I. Historical Review of Objective Personality Testing: the Search for Objectivity. In Bass, B. M. and Berg, I. S. *Objective Approaches to Personality Assessment*. Princeton: Van Nostrand, 1959.

Educational and Psychological Measurement
1966, 26, 691-700.

THE RELATIONSHIP OF SELF REPORT TO INFERRED SELF CONCEPT[1]

JAMES PARKER
Board of Public Instruction
Pinellas County, Florida

Both the self report and the self concept are currently among the most popular subjects for psychological research. Wylie (1961) for example, has reviewed 463 such articles. Many of the studies reported by Wylie as well as many more reported more recently have utilized the self report as a direct measure of the self concept. This use of the two terms as though they were synonymous has been vigorously challenged by Combs and his associates (Combs and Snygg 1959; Combs and Soper 1957; Combs, Soper, and Courson, 1963). They point out that the self concept is an internal organization of the individual's perceptions about himself, whereas, the self report is a behavior representing what the individual is willing and able to say about himself when he is asked to declare his position. They insist that the self concept must be approached through some form of inference based on the individual's behavior. This research was designed to explore the relationship of self report to inferred self concept in the hope of shedding some light upon this confusion.

Problem

Theoretically, several variables may operate to produce an undetermined amount of distortion in the self report when it is used to measure the self concept. Combs and Snygg (1959) and Combs

[1] Based on a dissertation submitted to the University of Florida. The author expresses his gratitude to Arthur W. Combs for his generous assistance in directing the study.

and Soper (1957) list factors which may influence an individual's self report:

1. The clarity of the subject's awareness.
2. The availability of adequate symbols of expression.
3. The willingness of the subject to cooperate.
4. The individual's feeling of personal adequacy.
5. The individual's feeling of freedom from threat.
6. The social expectancy.

If these factors do, in fact, interfere with the reliability of the self report, self concept study with such instruments will produce questionable conclusions.

According to theory, the self concept is the individual's personal organization of perceptions about self. Since this organization affects the individual's behavior the self concept can be studied by reversing this process and inferring the nature of the self from behavior. The method used to elicit behavior should minimize the possibility of influence by factors which are claimed to affect the self report.

The general aim of this study was to examine differences between the self reports made by children and their self concepts as inferred from their behavior. These hypotheses were formulated:

1. The self report and the inferred self concept will show low correlation when each is obtained in a way that protects the anonymity of the subjects.
2. Unsigned self reports will not correlate highly with self reports which the subjects are required to sign.
3. There will be high correlation between inferred self concepts when subjects believe that their responses are held in confidence and inferred self concepts obtained when subjects know that their responses will not be held in confidence.
4. The correlation between signed self reports and inferred self concepts obtained when subjects know that their responses are not confidential will be lower than the correlation between unsigned self reports and inferred self concepts made when the subjects believe that their responses are confidential.
5. The teacher, who is well acquainted with the subjects, will tend to more readily recognize individuals from inferences made by the experimenter than from the subjects own self reports.

Method

Subjects

Subjects cooperating in this study were thirty sixth grade children enrolled in a public school in an urban area.

Research Instruments

The study was limited to five areas of the self concept which could be expected to be expressed in behavior in the school environment: the self, the self in relation to others, the self as achieving, the self in school, and the physical self. Two research instruments were developed as means of testing the hypothesis.

Self Report-Inferred Self Concept Scale

A rating scale was devised which would serve as a self report instrument filled out by the children as well as a score sheet for recording the inferred self concept from the picture story test. It was a five-point scale comprised of six items for each of the five areas of the self concept which were being investigated. The items were arranged in pairs of positive and negative statements at either end of a continuum. Some of the items were reversed so that positive and negative statements did not always run in the same direction. The rating scale, then, had thirty items.

Items	
1. I'm good in school work	I'm not good in school work
2. Mostly I have good ideas	My ideas are poor
3. I'm a worthwhile person	I'm not a worthwhile person
4. I'm pretty strong	I'm not too strong
5. Most people trust me	Most people don't trust me
6. Teachers like me pretty well	Teachers don't like me too much
7. I can do most things well	I do very few things well
8. I'm a happy person	I'm an unhappy person
9. I'm healthy	I'm not too healthy
10. I'm popular	I'm not too popular
11. I'm a good reader	I'm not a good reader
12. I'm a hard worker	I'm not a hard worker
13. I'm very shy	I'm not shy
14. I don't get tired quickly	I get tired quickly
15. Other people find me interesting	I'm not too interesting to others
16. I work well with others in school	I don't work well with others in school
17. I'm pretty brave	I'm not too brave

18. I'm pretty smart	I'm not very smart
19. I'm not tall enough	I'm tall enough
20. Most people are fair with me	Most people are unfair with me
21. I don't do so well in class discussions	I do well in class discussions
22. I handle most of my problems well	I can't handle my problems very well
23. I'm a helpful person	I'm not too helpful
24. I'm good looking	I'm not too good looking
25. Most people are hard for me to get along with	Most people are easy for me to get along with
26. I'm mostly happy in school	I'm mostly unhappy in school
27. I can usually finish what I start	I never finish most things
28. I'm proud of me	I'm not too proud of me
29. I handle my body well in sports and games	I don't handle my body well in sports and games
30. I'm not often sorry for others	I'm often sorry for others

Picture Story Test

A picture story test was devised as a technique for eliciting behavior upon which the author could base inferences about the self concepts of the subjects. This test was comprised of eleven simple black and white drawings depicting school scenes designed for individual administration. The self concept inferred from each subject's responses was recorded on a fresh self report-inferred self concept scale. This allowed comparison of each subject's own self report with his self concept as inferred from his picture story responses under the various conditions of the study.

Procedure

Phase I

The project was introduced to the class of thirty sixth graders as an effort to learn how boys and girls regard themselves in school and how imaginative they are. It was explained that the experiment was expected to produce information which would be useful in helping children learn better.

For the initial administration of the self report, the children were promised anonymity. This was accomplished by not requiring the children to sign their self reports, although each was coded so that identification was, in fact, possible.

The picture story test was presented to each child in a manner

which implied confidentiality. The children were told not to tell their names to the examiner and they were assured that the stories they supplied would not be revealed to anyone. Actually a system was worked out with the teacher whereby identification could be made.

After both the self report and the picture story test had been administered, the author told the class that the experiment was finished and it had furnished valuable information about children.

Phase II

The object of the second phase of the experiment was to observe the influence of social expectancy on both the self report and the inferred self concept.

The teacher informed the class that she knew the author had learned much from his experience with them. She recognized that she could not have access to the information they had given because it was confidential. However, she thought that similar data about individuals would be helpful to her as a teacher. The children agreed to take the tests again.

Before administering the self report and the picture story test a second time, much stress was placed on the idea that information gained would be given to the teacher. It was, therefore, of great importance that the children sign their self reports and that they exercise great care in completing them. Also each person was informed when he appeared for the picture story test that his story material would be interpreted to the teacher. Thus, it was assumed, the subjects were made highly aware of the social expectancy factor.

Phase III

In the third phase of the experiment it was reasoned that, if the inferred self concept was a more accurate reflection of the individual, a teacher who knew the subjects well would see ratings based on inference as more descriptive of the person than his own self report. Data were used from the second phase of the experiment when social expectancy was stressed. This choice was based on the idea that in the actual school situation information given by pupils on personality tests would probably be influenced to some degree by social expectancy. Therefore, the ratings from phase two

more closely approximated what would be yielded in an actual attempts to learn about children in the classroom.

The results of the signed self reports were reproduced on fresh rating scales by use of a rubber stamp. The results of the second administration of the picture story test were reproduced in the same way. This made it impossible for the teacher to know whether she was examining a subject's self report or his inferred self concept. The pile of rating scales containing paired reproductions of the self report and the inferred self concept was given to the teacher. She was asked to select the one rating scale for each subject which she believed to be the better description of him.

Results

To determine the relationship of unsigned self reports to the self concept inferred from picture stories obtained under anonymous conditions, Pearson correlations were computed for each of the five aspects of self under investigation. Table 1 summarizes the results.

TABLE 1

Correlation of Unsigned Self Reports and Inferred Self Concepts When Stories Were Held in Confidence. N = 30

Aspect of Self Concept	r
1. The self	.213
2. The self in relation to others	.140
3. The self as achieving	.166
4. The self in school	.312
5. The physical self	.359
Average Correlation	.245
.05 level of significance	.360

There were generally low correlations with none exceeding the .360 value required for the .05 level of significance. The average correlation was ascertained by transforming each correlation to Fisher's z value, averaging, and converting back to a single r. The result was .245, a value significantly different from zero at the .01 level of significance. This suggests that there was low correlation between the self report and the inferred self concept even under conditions which would seem to minimize threat to the individual.

Correlations were run between the signed and unsigned self reports. The results are presented in Table 2. It is apparent that

there was a rather high degree of agreement between the self reports made under anonymous conditions and those made when the subjects were not promised anonymity. The implication is that the introduction of the social expectancy factor did not greatly influence the subject's responses on the self report.

TABLE 2

Correlation of Unsigned and Signed Self Reports. N = 30

Aspect of Self Concept	r
1. The self	.804
2. The self in relation to others	.735
3. The self as achieving	.718
4. The self in school	.652
5. The physical self	.763
Average Correlation	.740
.01 level of significance	.463

When correlations were computed for self concepts inferred from stories told under conditions of confidentiality with those told when confidentiality was not guaranteed, the relationships were fairly high. Thus, the inferred self concept did not reflect change which could be attributed to the introduction of social expectancy. The data is summarized in Table 3.

TABLE 3

Correlation of Inferred Self Concept when Responses Were Held in Confidence and when Responses Were Not Held in Confidence. N = 30

Aspect of Self Concept	r
1. The self	.775
2. The self in relation to others	.791
3. The self as achieving	.605
4. The self in school	.685
5. The physical self	.564
Average Correlation	.695
.01 level of significance	.463

It was expected that correlations between signed self reports and self concepts inferred from picture tories not held confidential would be lower than correlations between unsigned self reports and inferred self concepts from picture stories which were held confidential. Table 4 presents the data.

The r's under confidential conditions are all positive and above .140; whereas four of the 5's under non-confidential conditions are negative. This might lead one, on cursory examination, to conclude

TABLE 4

Correlation of Self Report and Inferred Self Concept when Confidentiality Was Granted and when it Was Not. N = 30

Aspect of Self Concept	r Under Confidential Conditions	r Under Nonconfidential Conditions
1. The self	.213	−.099
2. The self in relation to others	.140	.034
3. The self as achieving	.166	−.037
4. The self in school	.312	−.019
5. The physical self	.359	−.117
Average Correlation	.245	−.046

that there were pronounced differences between correlations for each individual aspect of the self concept. However, when these differences were tested there were no instances in which the critical ratio was as great as .196, the value needed for the .05 level of significance. When the average correlation for all five aspects of the self concept was calculated, through the r-to-z transformation, the difference between the two r's (.245 and −.046) was significant at the .01 level. On the basis of these results, then, it is concluded that there was a significant difference between the self report and the inferred self concept when confidentiality was not promised. In addition, the average r under non-confidential conditions (−.046) was not significantly different from zero, hence it may also be said that the two sets of results do not have great relationship.

It was anticipated that the teacher of the children who acted as subjects would more readily recognize individuals from their self concept as inferred by a trained observer than from their own self report. The significance of the difference between observed and expected probability was tested (Walker and Lev, 1953).

The teacher's task necessitated thirty choices ($N = 30$). Theoretically it could be expected that, by chance, she would select 50 per cent of the inferred self concepts and 50 per cent of the self reports. Actually she chose the inferred self concept over the self report 70 per cent of the time. A test of the significance of the difference between observed and expected p yielded a value of 2.19, significant between the .01 and .02 level. It is apparent that the teacher tended to view the inferred self concept as a more accurate description of her pupils.

Conclusions

The results of this research furnish the basis for several conclusions.

1. The data produced by the study lend support to the claim that the self report and inferred self concept do not furnish the same insight into the personality of individuals. Even when conditions are provided which should allow the greatest freedom from threat and maximum opportunity for frankness the two methods of observing people show little similarity. Therefore, the assertion that the self report and the self concept are different is substantiated by this research.

2. While this research found the self report to be consistent even though social expectancy was introduced, there is no basis for concluding that the self report accurately reflects children's perceptions of self. The fact that some of the children's experiences in school did not seem conducive to the development of the essentially positive views of self they reported leads to hesitancy in accepting the self report as a measure of the self concept.

3. As anticipated, the inferred self concept was not greatly influenced by social expectancy. There is some reason to contend that the inferred self concept provided more realistic insight into individual perceptions. First, the researcher's ratings were somewhat less positive than were the children's self reports and in several cases the self concept ratings were more in line with what would be expected, according to perceptual theory, in the development of self perceptions in light of individual experiences as revealed by school records. Further, the teacher who knew the children through daily experience tended to support the ratings made from the inferred self concept as generally in agreement with her own view of individuals. The conclusion is that the inferred self concept represents a more accurate and realistic appraisal of children's perceptions of self than do self reports.

4. Correlations between self report and inferred self concept diminished when social expectancy was emphasized. It is apparent that external changes can cause differences in statistical relationship between the self report and inferred self concept. However, the data does not lend itself to a clearcut conclusion about the reasons for these differences.

5. The teacher recognized individual children more often from ratings made by the researcher from inference than from the children's own self ratings. This suggests that two adults having similar basis for understanding children, but using quite different methods of studying behavior, are likely to arrive at fairly similar impressions of children. A related conclusion is that the picture story approach offers a means of gaining information about children in a short time which is as good as knowledge accumulated over a longer time through informal association with children.

Summary

This study inquired into the relationship of the self report and the inferred self concept first under conditions calculated to reduce the effects of social expectancy, and again when social expectancy was emphasized. The results show that both the self report and the inferred self concept remained relatively consistent under the varying conditions. However, when the correlations between the self report and the inferred self concept before emphasizing social expectancy were compared with correlations between the self report and the inferred self concept after emphasizing that factor, there was a decrease in the statistical relationship. This suggests that some change did result from the operation of social expectancy.

A further test was made by asking the teacher of the pupils who acted as subjects to select blindly a self report or a rating based on the inferred self concept as more nearly agreeing with her own impressions of each child. The teacher chose the inferred self concept rating in 70 per cent of the cases.

REFERENCES

Combs, Arthur W. and Snygg, Donald. *Individual Behavior*. New York: Harper, 1959.

Combs, Arthur W. and Soper, Daniel W. "The Self, Its Derivate Terms and Research." *Journal of Individual Psychology*, 1957, 13, 134–145.

Combs, Arthur W. Soper, Daniel W., and Courson, Clifford C. "The Relationship of Self Concept to Self Report." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 23, 1963, 493–500.

Walker, Helen and Lev, J. *Statistical Inference*. New York: Holt, 1953.

Wylie, Ruth C. *The Self-Concept: A Critical Survey of Pertinent Research*. Lincoln: University of Nebraska Press, 1961.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 701-702.

ELECTRONIC COMPUTER PROGRAM AND ACCOUNTING MACHINE PROCEDURES

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

IMPORTANT NOTICE TO AUTHORS

In view of the tremendous advances that have been made in the adaptation of electronic computers and accounting machines to the processing of statistical data, sections of the Spring and Autumn issues of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT are devoted to the publication of such programs as are appropriate to psychometric procedures. Programs relevant to such problem areas as factor analysis, item analysis, multiple regression procedures, the estimation of the reliability and validity of tests, pattern and profile analysis, the analysis of variance and covariance, discriminant analysis, and test scoring will be considered. Customarily a program should be expected not to exceed six or eight printed pages. Manuscripts of four or fewer printed pages are preferred. Each manuscript will be carefully reviewed as to its suitability and accuracy of content. In some instances an accepted paper may be returned to the author for possible revisions or shortening.

Due to the financial pressure of rising printing costs, the new rate for publication of articles in this section, beginning with the first issue for the year of 1967, will be twenty-five dollars per page. The extra cost of the composition of tables and formulas will be added to the basic rate.

Manuscripts received up to November first will be considered for the Spring issue; manuscripts received between then and May first will be considered for the Autumn issue.

All correspondence and manuscripts should be directed to:

William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California 93106

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 703-705.

A GENERALIZED ONE-WAY ANALYSIS OF VARIANCE PROGRAM IN FORTRAN-II[1]

ALLAN J. NASH
Florida Atlantic University

THIS program was developed to provide intermediate computer systems with maximum flexibility in the analysis of variance for single factor experiments. Its primary feature is a sense switch option to analyze differences among k treatment levels according to a model which assumes correlated observations (repeated measures or matched groups designs) or according to a model which assumes non-correlated observations (randomized groups designs). The statistical models underlying these two basic analyses have been described by Winer (1962).

The program was initially written for the IBM 1620; but since it requires no special equipment or subroutines, it can be used on most systems with a FORTRAN-II compiler. The program itself requires approximately 8K core storage capacity. Since the complete data matrix is needed only when the correlated data option is exercised, it is possible to conserve core storage during a non-correlated data analysis by cycling each observation individually through the preliminary summations with little cost of machine time. This feature, in addition to the program's ability to process unequal sample sizes, permits the analysis of any single-factor experiment which assumes uncorrelated observations, regardless of the design dimensions or additional core storage capacity. When correlated observations are analyzed, the complete data matrix (k treatments $\times$ n subjects) is required and the remaining core storage will determine the maximum permissible dimensions of the

[1] The author is indebted to the Computing Center of Temple University for assistance in developing and testing this program.

design. The formula, 10 $(kn + n + 3k)$, gives the number of additional cores needed for processing correlated data.

Dimension statements must be changed in the program for different values of k and n. It is also necessary to modify the data input formats if the card-column fields change from one analysis to another. The absence of a variable format capability thus necessitates recompilation for analyses with differing design or data format characteristics. This disadvantage is balanced by the elimination of the special subroutines and equipment needed to achieve variable format capability.

Input

I. Correlated Observations (sense switch OFF)
  1. Source deck with dimension and data format statements appropriate to the specific analysis.
  2. A parameter card defining values for k (number of treatments or classifications), n (number of subjects or blocks of correlated observations), sample sizes for the k treatments (must be constant), and the total number of observations.
  3. The n data cards, in any order, each containing k correlated observations.

II. Uncorrelated Observations (sense switch ON)
  1. Source deck with dimension and data format statements appropriate to the specific analysis.
  2. A parameter card defining values for k (number of treatments or classifications), n (a control constant set equal to 2), sample sizes for the k treatments (need not be constant), and the total number of observations.
  3. The data cards, ordered by treatment level, each containing one observation.

Output

In addition to treatment means, standard deviations, and sample sizes, the following output is given:

I. Correlated Observations
  1. Mean squares and degrees of freedom for between-subjects and within-subjects variation. This latter source of variation is further broken down to give Mean squares and degrees of freedom for treatment variation and the residual error term.

2. F-ratio: $F = MS_{\text{treatments}}/MS_{\text{residual}}$

II. Uncorrelated Observations

1. Mean squares and degrees of freedom for between treatment variation and within treatment (error) variation.
2. F-ratio: $F = MS_{\text{treatments}}/MS_{\text{error}}$

REFERENCE

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill Book Co., 1962.

Educational and Psychological Measurement
1966, 26, 707-708.

A NOTE ON A MODIFICATION OF COOLEY AND LOHNES' CLASSIFICATION PROGRAM[1]

GEORGE H. DUNTEMAN
College of Health Related Professions
University of Florida

Although Cooley and Lohnes' classification program (Cooley and Lohnes, 1962) was designed primarily to compute classification equations in the reduced discriminant function space, their program also contains provisions for obtaining quadratic (non-linear) discriminant equations in the original test space. The output for the classification program includes the group membership predictions based upon two basic quadratic discriminant equations. The first quadratic form is $(X - \overline{X}_i)' D_i^{-1} (X - \overline{X}_i)$ where $(X - \overline{X}_i)$ is a vector of deviation scores from the test means of group i and D_i^{-1} is the inverse of the dispersion (variance — covariance) matrix for group i. The second quadratic form results in a probability statement of group membership and is proportional to $(X - \overline{X}_i)' D_i^{-1} (X - \overline{X}_i) + \log_e |D_i| - 2 \log_e P_i$ where $|D_i|$ is the determinant of the dispersion matrix for group i and P_i is the *a priori* probability of being a member of group i.

The second equation is optimal in the sense of minimizing misclassification. However, it assumes that the *a priori* probabilities of group membership are known. In most instances involving psychological and educational research, the number of people included in the analysis for each group is arbitrary (i.e., the number and proportions of people included in the various groups is a function of availability) and consequently the *a priori* probabilities are mean-

[1] This research was supported in part by a research grant, #RD-1127, from the Vocational Rehabilitation Administration, Department of Health, Education, and Welfare, Washington, D. C.

ingless and their inclusion in the classification equation distorts the picture of classification efficiency. On the other hand, the first equation does not utilize all the available information, since it assumes that the determinants of the group dispersion matrices are equal when in fact the ellipsoids in the multivariate space may not be of equal size.

A more useful and general quadratic discriminator, utilizing differences among the dispersions of various groups but assuming equal *a priori* probabilities of group membership is $(X - \overline{X}_i)' D_i^{-1} (X - \overline{X}_i) + \log_e |D_i|$. This quadratic form degenerates to the linear form if the dispersion matrices are equal. This last equation may be easily obtained from Cooley and Lohnes' program since a subroutine called by the program has already calculated $|D_i|$ for inclusion in the Bayesian probability equation. An addition of a few Fortran statements easily results in the computation of $\log_e |D|$ and its addition to the first equation already computed by the program.

REFERENCE

Cooley, W. and Lohnes, P. *Multivariate Procedures for the Behavioral Sciences*, New York: John Wiley & Sons, 1962.

Educational and Psychological Measurement
1966, 26, 709-710.

A FORTRAN GENERATOR OF TEST SCORES ACCORDING TO A PREDETERMINED FACTOR PATTERN

SIDNEY M. ROSENBLATT[1]
Teachers College, Columbia University

An interesting method of testing the truth of factor analysis is to factor analyse test scores which already have definite factors built into them. This method was first used by McNemar (1941) to test the sampling error of factor loadings. The test scores are determined by the linear factor model.

$$X = f_1\ (R_1) + f_2(R_2) + f_3(R_3) + \ldots + f_n(R_n) + f_s(R_s) + f_{e1}(R_{e1}).$$

$f_1, f_2, \ldots, f_n$ are the factor weights in the factor pattern.

f_s is a term put in to account for specific variance.

f_e is a term put into the model to account for error variance.

$R_1, R_2, \ldots, R_n$ are normalized random number weights generated by the computer.

$X_1, X_2, X_3, \ldots, X_n$ are test scores.

The program provides a punched card output of the test scores to facilitate factor analysis of the test scores. The input of the program is a control card describing the number of students and the number of tests for each student. For each test used in the battery a data card is prepared giving the factor weights, the error portion of the variance, and the specific portion of the variance.

This program used a random number generator subroutine described in Gruenberger and McCracken (1963) and a subroutine which transforms the random number into a normalized random

[1] This program was developed by the author in cooperation with Mr. Joel Herbsman, formerly of the Teachers College Computer Center and now of the University of Buffalo Computer Center.

number as determined by the equation for the normal probability distribution.

The author generated three samples of twenty test batteries on the IBM 1620 II, through using this program. The test batteries, which contained from ten to twenty tests, were generated for 100 students. The running time to generate twenty test batteries, five containing ten tests, five containing twelve tests, five containing sixteen tests, and five containing twenty tests, for 100 students was approximately one hour on the IBM 1620 II.

A program deck is available from the Teachers College, Columbia University Computer Center, filed under the author's name.

REFERENCES

Gruenberg, F. J. and McCracken, D. D. *Introduction to Electronic Computers: Program Solving with the I.B.M. 1620.* New York: John Wiley & Sons, 1963.

McNemar, Q. On the Sampling Error of Factor Loadings. *Psychometrika,* 1941, 6, 141–152.

Educational and Psychological Measurement
1966, 26, 711-713.

A FORTRAN-IV PSYCHOLOGICAL TEST-SCORING PROGRAM

SIDNEY R. ADELMAN and WILLIAM H. McWHINNEY
Graduate School of Business Administration
University of California, Los Angeles

Purpose

This program analyzes multiple-response questionnaires in which the test score is a series of individual factor scores. It takes either a two-, three- or a five response questionnaire, and computes up to fifteen factor totals. A total factor score is the sum of the responses of the questions associated with or keyed for the factor. For every case, the program determines: (1) total factor scores, (2) the standard deviation for each score, (3) the percentage of the total scores or the total number of items on each factor scale that are answered in the keyed direction and (4) the percentile value corresponding to the total score on each factor. Norms are developed for the entire sample, and if desired, an analysis of the norms by ages is also available. It has been used to score the Buhler-Coleman Life Goals Inventory (Buhler, 1963) and the Personal Orientation Inventory (Shostrom, 1963).

The user has the option of computing Cronbach and Gleser's (1953) D-square in which the sum of the squared differences is determined for each pair of cases. This matrix can be punched in a form which is accepted by Ward and Hook's (1963) Hierarchial Grouping Program. Total factor scores as well as the corresponding percentage scores (mentioned in (1) and (3) respectively in the preceding paragraph) can also be punched at the user's discretion.

Program Specifications

Control Cards

A variable format statement is used to specify the form of each case input. The first control card specifies the number of cases, of questions, and of factors. This card is also used to specify the use of the options which are available. The second control card indicates the number of questions in each factor. The next set of cards indicates the question numbers for each factor (one card for each factor). If an inverse rating is to be given to a particular question, the question number is preceded by a minus sign. The next set of cards contains previously established norms.

These norms will be compared against each case's total factor score (mentioned in (1) in the first paragraph). If these norms have not yet been determined, or are not desired, a blank card for each factor must still be included.

Data Input

Case input may be in the form of punched tabulating cards or magnetic tape. Each case is identified by a three- or four-digit code number, a required indication of sex, an optional age designation, followed by the question scores. The program allows either two (1 or 2), three (1, 2 or 3), or five (1, 2, 3, 4 or 5) responses. It checks for invalid responses and when such are found, the code number as well as the offending question number is indicated. When no response has been made, either this outcome can be considered to be invalid, or a three is assigned in the case of five possible responses; a two is assigned when three responses are possible; and in the two response situation, a one alternating with a two is assigned.

Capacity

The program can now accommodate 400 cases, 160 questions and 15 factors. By manipulating the DIMENSION statement, modifications can be made to handle programs with a variety of parameters.

Running Time

Running time depends on the number of options employed. When the scoring routine alone is used, the IBM 7094 uses on the order of one second for each case.

Availability

The program was written in FORTRAN-IV. It was compiled and "debugged" at the Western Data Processing Center at UCLA. Source decks, object decks, and appropriate documentation are available from the senior author.

Summary

This paper has described a general purpose scoring program geared predominantly for questionnaires. Its major attributes are its flexibility and ease of modification.

REFERENCES

Buhler, Charlotte. Questionnaire on Goals and Fulfillments. *Journal of Humanistic Psychlogy*, 1963, 3, 28–34.

Cronbach, L. J. and Gleser, Goldine C. Assessing Similarity between Profiles. *Psychological Bulletin*, 1953, 50, 457–459.

Shostrom, E. L. *Personal Orientation Inventory*, Educational and Industrial Testing Service, 1963.

Ward, J. H., Jr. and Hook, Marion E. Application of an Hierarchial Grouping Procedure to a Problem of Grouping Profiles. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 69–81.

Educational and Psychological Measurement
1966, 26, 715-717.

SCORING AND ANALYZING TEACHER-MADE TESTS WITH AN IBM 1620

J. MICHAEL O'MALLEY
San Jose City College
AND
CURT STAFFORD
San Jose State College

A recent article (Stafford and Bianchini, 1963) described a program by which teacher-made objective tests could be scored by an IBM 1620 after students had recorded answers on a special mark-sense answer card. The program was developed for the purpose of scoring and analyzing classroom tests in a fashion which previously had been restricted to standardized tests. The output included scores on the total test; frequency distribution and T scores; the frequency and per cent with which each item alternative was selected (choice analysis); and a statistical analysis consisting of a count of the number of tests, the mean, standard deviation, standard error of measurement, and a KR-21 reliability estimate. Limitations imposed were: one response per item, a maximum of 486 items and 999 examinees. The interest shown by local faculty and by faculty at other institutions indicates that the system has contributed significantly to the educational process.

Two Revisions in Previous Program

The purpose of this article is to describe two separate revisions of the first scoring program and to project possible future changes.

The revision to be described first is known as SSS VI (Sure Swift Scoring, Revision VI). The major modifications are these: (a) provision of biserial correlations between item response and total score; (b) provision of up to four subscores, each permitted

a different scoring formula; (c) calculation of a KR-20 estimate of reliability, a more accurate estimate (Cronbach, 1962) of test reliability; (d) batch processing of consecutive sets of tests; (e) formating compatible with 80/80 listing; and (f) suppression of any combination of the four major program functions, (scoring, statistics, frequency distribution and *T* scores, and choice analysis with biserial correlations) by digit punch in the parameter card. There have been modifications of a rather minor nature in the format of the parameter card and the output, but the new program still scores up to 486 items for 999 examinees.

The second revision to be described, SSS VII, was developed because of requests for Flanagan correlations by faculty not familiar with biserials. Initially, Flanagan correlations based on proportions correct in the upper and lower 27 per cent of the score distribution were provided through two separate programs, one for scoring and one for determining Flanagan coefficients. A two-pass procedure obviously was both cumbersome and time consuming. SSS VII scores tests and computes correlations in a single pass for a maximum of 150 examinees, 162 items, and 10 sets in a batch. These more restrictive limitations were imposed by the necessity of storing both a large table of Flanagan correlations and the item responses of every examinee. In all other respects, however, SSS VII has the same flexibility as does SSS VI.

Projected Modifications

Programs developed to date have been for an IBM 1620 with 40K core storage and with special features of indirect address and automatic divide. The recent addition of the 1311 disk drive greatly increases readily accessible storage. Immediate modifications will be to expand the relatively narrow restrictions on the SSS VII (Flanagan) program, and to include *T* scores with the individual examinee data. It would be highly desirable to add student name to the output, a feature which some local junior colleges already have incorporated into their systems. Some faculty members have requested individual examinee item error lists and accumulation of test scores over the semester; these will be investigated.

The Testing Office, San Jose State College, solicits information from other classroom test scoring service agencies, and conversely

is willing to share information and source or object decks from its library.

REFERENCES

Cronbach, L. J. and Azuma, Hiroshi. "Internal-consistency Reliability Formulas Applied to Randomly-sampled Single Factor Tests: An Empirical Comparison." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 645–665.

Stafford, C. and Bianchini, J. "Scoring Teacher Made Tests with an IBM 1620." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 581–586.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 719-723.

RELIABILITY AND VALIDITY OF THE DIGITEK OPTICAL SCANNER[1] IN TEST SCORING OPERATIONS

RICHARD E. SPENCER
University of Illinois

WITH the introduction of optical scanning equipment to the field of testing, it seems proper to determine degrees of reliability and validity of these relatively new and unique machine systems. Reliability, in normal measurement terms, can be defined as relating to the degrees of consistency of the internal aspects of the particular system. Reliability of a machine system can be computed analogously, as follows:

1. Consistency between the printed and the punched test score.

In image mode, the Digitek scores (i.e., counts) the number of correct marks, prints the test scores on the answer sheet, and punches a card with the score and items through the punch unit. The degree to which the printed score (on the answer sheet) agrees with the punched score on the card is an indication of the internal reliability of the machine system.

2. Consistency between the item read unit and the item punch unit.

A second method to determine reliability is through the use of the item punches in the card output. The score printed on the answer sheet is determined by the optical read system in picking up the marks on the answer sheet. These marks are transferred into storage in a direct "image" of the answer sheets, and the memory is "dumped" into the card output. If one were to read these punched output cards as item responses, the test score developed by reading the punches should equal the test score punched in that same card. The degree to which such scores are equal can indicate the reliability of the machine system.

[1] Digitek Corporation, Fairless Hills, Pennsylvania.

3. Consistency of the print unit.

A third measure of internal reliability of the system can be determined by the degree to which the Digitek yields the same results a second, third, or fourth time.

The validity of the machine system, again defined in the measurement sense, can be defined only with the use of some outside criterion. In the case of a test scorer, validity can be determined by comparing the results obtained through the use of the Digitek with results determined by some other means. In this instance, hand scoring can be used to check the validity of the Digitek test scoring.

These reliability and validity definitions can be summated as follows:

Reliability #1: Score printed on answer sheet compared to score punched on output card.

Reliability #2: Score punched on output card compared to score derived from item "image" punches.

Reliability #3: Score of first machine pass compared to score on successive machine passes.

Validity #1: Digitek derived scores compared to hand derived scores.

Procedure

Computation of reliabilities number 1 and 2 was performed on tests used in undergraduate foreign language programs, Spring semester 1964–65. Digitek answer sheets were used (University of Illinois Answer Sheet MLA). The tests were the Modern Language Association, Cooperative Foreign Language Tests in French Reading and Listening Comprehension, (Form LB and MB). The experimental sample was composed of students enrolled in French 101, 102, 103, and 104 during the Spring semester at the University of Illinois. The following tabulation indicates the number of students in each sample:

Course	Test	No. of Items	No. of Scores	No. of Subjects
French 101	Reading LB	50	1	174
	Listening LB	45	1	164
French 102	Reading LB	50	1	422
	Listening LB	45	1	417
French 103	Reading MB	50	1	149
	Listening MB	40	1	151
French 104	Reading MB	50	1	491
	Listening MB	40	1	491

There were two answer sheets per student; one for the reading test and one for the listening tests. The answer sheets were scored on the Digitek, and the test score printed on the answer sheet, punched on a card output, and the item responses were also punched out on the output card. A listing was made of the total score by student as punched in the output card, and clerks compared the listing with the printed score on the answer sheet. It was intended that a list of the two scores would be punched, the data submitted to a computer, and a correlation computed. However, since no differences were found, the correlation was not computed. The results indicated a perfect relationship between the two sets of cards (2444 score sets). Reliability #1 was 1.00.

Reliability #2 was determined by submitting the Digitek item score cards to a computer item analyses/test scoring program with the appropriate key for each test. Four sample sets were developed. Each punched test score was compared with the computer test score, and a correlation determined. The results were as follows:

1. French MB Listening, $N = 642$, $r = 1.00$
2. French MB Reading, $N = 635$, $r = 1.00$
3. French LB Listening, $N = 571$, $r = 1.00$
4. French LB Reading, $N = 596$, $r = 1.00$

These results indicate a perfect relationship between the score developed by the optical read system, and the score determined by the punched item output. Reliability #2 was 1.00.

Reliability #3 was determined on a different sample of students. A class of elementary biology (Biology 100) from the Chicago Campus of the University of Illinois comprised the sample, containing 1123 students. The tests were packaged and sealed at the site of test administration, and opened upon receipt at the Digitek. The tests were in no way specially prepared for the Digitek run. The students used regular lead pencils provided by themselves. Three forms of the test were used, Form A, B, and C, each with 20 items. Each form was scored separately with the Digitek "select" switch on, and the "multiple response" switch turned to select. Each test paper was processed three times. The following results were noted:

Form A, $N = 284$

One paper selected out all three runs, because of one question being marked twice, and one question being marked three times.

Six differences were noted as follows:

1. 15, 14, 15
2. 8, 12, 12
3. 4, 6, 6
4. 6, 12, 12
5. 5, 8, 8
6. 14, 16, 16

Form B, $N = 418$

Two papers were selected out all three times due to unclear erasures.

No differences in test scores were noted.

Form C, $N = 421$

Eight papers were selected out all three times because of double and triple marks.

One difference was noted as follows:

1. 14, 14, 13

Out of 284 papers in Form A, 6 differences occurred
Out of 418 papers in Form B, no differences occurred
Out of 421 papers in Form C, 1 difference occurred

Out of 1123 papers, 116 were scored consistently three times. Seven papers had a difference. The degree of consistency or reliability can be determined by the proportion of equal scores on all three runs; 99.38 per cent of the scores were equal all three times.

The validity of the system was checked on a second test given to the Biology 100 students from Chicago. On this separate occasion, 1086 tests were administered; again, three forms were used. Each test contained 20 test items. The tests were hand scored in order to obtain a validation criterion. Where differences occurred, the test was hand scored a second time. The following results were noted:

Form A; 1 error of 5 points
Form B; 1 error of 1 point
Form C; 3 errors of 1 point

Five error papers were discovered out of 1086 papers, an overall accuracy rate of 99.54 per cent.

In order to determine further the normal accuracy expected by other test scoring systems, one set of 203 test papers was hand scored prior to being submitted to the Digitek. After the Digitek

rescored the tests, a comparison was made, and again, differences were rescored by hand. Eleven differences were discovered. Upon rescoring, the Digitek scoring was found to be 100 per cent accurate, and the hand scoring (one time only) was found to be 96.11 per cent accurate.

Conclusion

The results obtained in this study of the reliability and validity of the Digitek Test Scoring System would indicate that values in excess of $r = .99$ can normally be expected.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 725-727.

A MACHINE SCORING ANSWER SHEET FORM FOR THE IBM 1231 OPTICAL SCANNER

JOHN A. FINGER, JR.
Rhode Island College

OPTICAL scanning scoring equipment has necessitated the redesign of answer sheets to meet the requirements of specific equipment. Unfortunately, there has been no standardization of equipment so that new forms are needed to suit the limitations of the scanners.

Both the Digitek and the 1230 IBM scanners are able to handle only one answer sheet at a time. Thus test batteries necessitating use of more than one answer sheet require either manual card collating or large size computers capable of collating magnetic tape or disk record. Many test batteries (for example, Houghton Mifflin's, 1230 answer sheets for the *Iowa Tests of Basic Skills* or the Psychological Corporation's *Differential Aptitude Tests*) have answer sheets requiring a student code number on every answer sheet. This requirement is time-consuming in test administration. Moreover, miscoding of a student number is a major problem in collating punch card output.

Test scoring equipment needs to be designed which can process all responses on a student's answer sheets at a single pass. Such a procedure can now be accomplished through utilizing the Digitek or the IBM 1231 optical scanner when it is connected to any computer capable of accepting its output.

Alterations are much needed in the design of both Digitek and IBM scanners. It should be possible with only minor design changes to provide multiple options with a scanner test scoring machine and computer as follows:

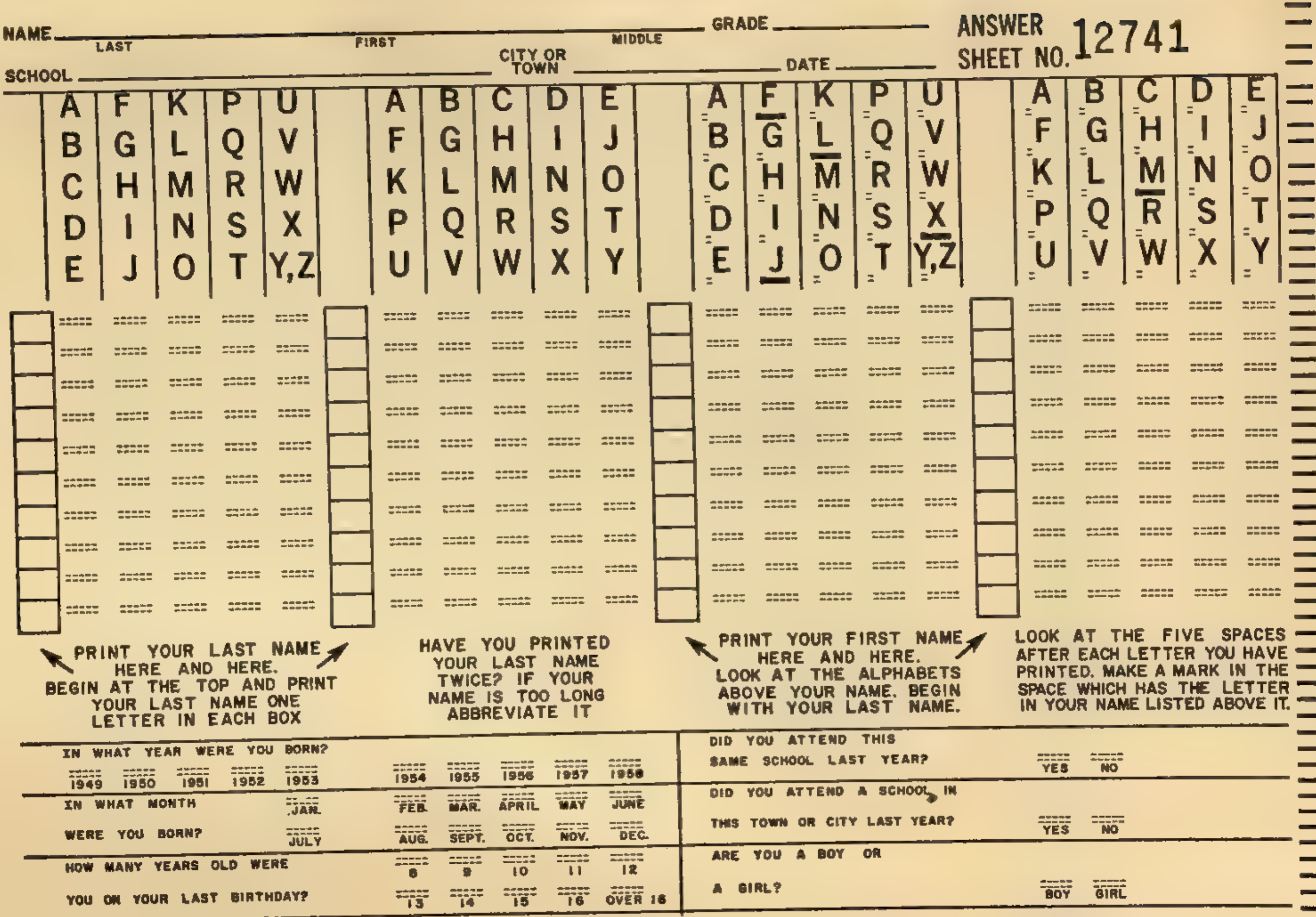
a. Separate test scoring without the computer.
b. Scanner scores tests and provides identification information and scores to the computer.
c. Scanner transmits bit by bit test information.

The Rhode Island State Testing Program utilizes answer sheets in which the student's name is gridded and then decoded by computer. Multiple answer sheets are precoded with a printed scannable number to assure that a student's complete record will be assembled. The computer (IBM 1440, 4K) which can accept scanner output for up to four answer sheets sequentially for one student, produces a single punch card containing the student name and scores.

The top portion of the first answer sheet of the packet of four (attached with a 3/4 inch left margin peforated stub) is shown as Table 1. Other answer sheets contain only answer responses except for the preprinted scannable number. The first digit of the scannable number identifies whether it is answer sheet one, two, three or four.

TABLE 1

Top Portion of the First Answer Sheet



EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 729-731.

DESIGNING AND PRINTING IBM 1230 OPTICAL MARK SCORING READER ANSWER SHEETS BY PHOTO-OFFSET

JOHN F. GUGEL
Hunter College

THE new IBM Optical Mark Scoring Reader can eliminate cumbersome hand key punching jobs. This machine when used with its key punch attachment (the 534 Key Punch), is capable of quick and accurate automatic transfer of information from special answer/response sheets to IBM cards. These answer/response sheets are coded with ordinary number 2 pencils. However, when they are designed and supplied commercially, the cost is high and delivery is slow.

Thanks to an attempt at printing the answer sheets by Max Ferder of Borough of Manhattan Community College, the writer has been able to develop a procedure for designing and printing them by photo-offset. Thus, not only costs can be reduced, but also individual code sheets can be provided for the specific demands of a single research project.

It is specified that the answer sheets must be in two colors—the response positions in a reflective ink and the timing check marks in the right margin in a non-reflective ink (black). However, two-color printing is not really necessary. Answer sheets can be made by ordinary photo-offset if the black ink is kept light. The surest way to guard against dark printing is to have the printer use the process newspapers employ to reduce the blackness of a picture. In this process, black areas are reduced to areas of small dots, barely visible to the naked eye. When this process is used on the response section of an answer sheet, the black ink is reduced to a shade of gray and the timing marks remain a heavy shade of black. A hue

of the response section called "fifty per cent of black" by the printer is quite adequate. A hue as light as "ten per cent of black" is possible.

Perfect operation of the machine's timing mark check system is crucial in assuring accurate data transfer; therefore, great care must be taken to keep the distance from the edge of the sheet to the timing marks at 11/32 of an inch. It is a good idea for the printer to keep the timing marks on the opposite side of any edge to be cut. This can help cut down on top to bottom right margin fluctuations which cause trouble. Also, the distance from the top edge of the paper to the first timing mark should be 1/4 of an inch. If these specifications are not adhered to, the responses may be missed, and the processed data will be inaccurate. The best way to assure accurate measurements is to use the Document Inspection Gage which is usually kept in the side shelf of the IBM 1230. This instrument shows what the answer sheet tolerances are.

With this photo-offset process, it is possible to construct response sheets to fit almost any test or questionnaire. For instance, in the preparation of a special response sheet for a demographic questionnaire administered to some twelve hundred Hunter College freshmen, a standard 0 to 9 response sheet was used. White paper was pasted over the areas to be deleted, and letters applicable to this particular questionnaire were typed on the sheet. Thus, for example, for sex of respondent two response positions were printed instead of the usual nine numbered positions (and 0). If, on the other hand, it is necessary to insert additional positions (within the limitations of machine design), response positions may be lined up through the use of a mimeoscope and then pasted or traced onto the answer sheet. A newer design procedure which the author has found very useful involves starting off with a standard 1,000 answer position response sheet. A white opaquing liquid (the new kind used to remove errors in typing) is applied to remove all unneeded lettering, numbering, and answer positions. New lettering and numbering is then typed onto the answer sheet or written onto it with a very fine point pen. Then the answer sheet is ready to be photographed for the photo-offset plate.

It is also possible to print the original questionnaire on a response sheet. However, subjects often make stray marks on the questionnaire which are then read by the machine as extra responses. Also

lettering or lines crossing a response position or positions can cause problems if in black ink. Such an answer sheet cannot be used in programming the machine, although it may be used as a data sheet if the marking allows programming of the machine not to consider the answer positions crossed or covered. Thus, separate answer sheets are advisable.

Many special purpose test and questionnaire answer sheets may be constructed with this photo-offset process. This process is not only a boon to the individual requirements of special research projects, but also an aid in reducing the cost of data processing.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 733-737.

AN IBM 1620 SPS COMPUTER PROGRAM FOR UNPACKING THE IBM 1230 SPECIAL CODE

JOHN F. GUGEL
Hunter College

WHEN a standard 1 to 5 or A to E choice answer sheet is used in the IBM 1230 Optical Mark Scoring Reader—534 Key Punch System, questions are dealt with two at a time. The IBM 503 answer sheet, for example, has the first and second questions double punched by the 534 into one card column. Questions three and four are double punched into the next column, and so forth. When the 534 drum card is programmed with a "3" punch for each of the card columns to be punched with information from the 1230, a special packed alphameric code results. Questions one and two become an alphameric character in one column; three and four become an alphameric character in the next column.

The Program

The purpose of this program is to translate the alphameric characters back into two columns. The type of answer sheet in which the choices are 1 to 5 or A to E then becomes represented by the punches 1 to 5 for each question. Question one is in one column; question two is in the next. Columns one to five of the input and output cards are reserved for an identification number. The identification number should be punched into the card by using the "2" punch on the 534 drum card program. Part or all of the number may be blank. If more than seventy-five questions are packed into the card, Program Switch 1 of the 1620 must be turned on. When Program Switch 1 is left off, only one unpacked card will be punched by the 1622 for each packed card fed into the 1620. Two output cards are always punched when Program Switch 1 is on

during the processing. Both blanks and "12" punches are translated as blanks when they occur.

1620 SPS Program

```
ØØ1         DORG 4Ø2
ØØ2F&ET     SF     INPUT+9,15Ø,29
ØØ3         AM     FSET+6,1,1Ø
ØØ4         SM     FSET+11,1,1Ø
ØØ5         BP     FSET
ØØ6         SF     NB-2
ØØ7         TF     BB+11,NB
ØØ8         TD     TWOB+11,NB
ØØ9         TD     FOURB+11,NB
Ø1Ø         TD     FIVEB+11,NB
Ø11         TD     ONEB+11,NB
Ø12         TD     B4+1Ø,NB
Ø13         TD     B5+1Ø,NB
Ø14         TD     B1+1Ø,NB
Ø15         TD     B2+1Ø,NB
Ø16         TD     B3+1Ø,NB
Ø17         TD     3B+11,NB
Ø18START    TFM    TRANS+11,INPUT1+1Ø
Ø19         TFM    CFW+6,INPUT1+9
Ø2Ø         TFM    TDW+6,INPUT1+1Ø
Ø21         RACDINPUT
Ø22         TR     INPUT1-1,INPUT-1
Ø23         BD     COL4,INPUT1+7
Ø24         TD     INPUT1+8,NB
Ø25COL4     BD     DIG4,INPUT1+5
Ø26         TD     INPUT1+7,NB
Ø27         B      COL3
Ø28DIG4     TD     INPUT1+7,INPUT1+6
Ø29COL3     BD     DIG3,INPUT1+3
Ø3Ø         TD     INPUT1+6,NB
Ø31         B      COL2
Ø32DIG3     TD     INPUT1+6,INPUT1+4
Ø33COL2     BD     DIG2,INPUT1+1
Ø34         TD     INPUT1+5,NB
Ø35         B      COL1
Ø36DIG2     TD     INPUT1+5,INPUT1+2
Ø37COL1     BD     DIG1,INPUT1-1
Ø38         TD     INPUT1+4,NB
Ø39         B      WORK
Ø4ØDIG1     TD     INPUT1+4,INPUT1
Ø41WORK     BC1    S1
Ø42         TFM    WOP,38,1Ø
Ø43         B      TRANS
Ø44S1       TFM    WOP,75,1Ø
Ø45TRANS    TF     WAREA,INPUT1+1Ø,7
Ø46         AM     WAREA,11,1Ø
Ø47         BP     BB
Ø48         AM     WAREA,3,1Ø
Ø49         BP     FOUR4
Ø5Ø         AM     WAREA,1,1Ø
Ø51         BP     FOUR5
Ø52         AM     WAREA,6,1Ø
```

```
BP    TWOB
AM    WAREA,1,1Ø
BP    FOUR1
AM    WAREA,2,1Ø
BP    FIVE4
AM    WAREA,1,1Ø
BP    FIVE5
AM    WAREA,9,1Ø
BP    THREE4
AM    WAREA,1,1Ø
BP    THREE5
AM    WAREA,17,1Ø
BP    THREE1
AM    WAREA,1,1Ø
BP    THREE2
AM    WAREA,1,1Ø
BP    TWO4
AM    WAREA,1,1Ø
BP    TWO5
AM    WAREA,1,1Ø
BP    TWO1
AM    WAREA,1,1Ø
BP    TWO2
AM    WAREA,1,1Ø
BP    TWO3
AM    WAREA,1,1Ø
BP    FOURB
AM    WAREA,1,1Ø
BP    THREE3
AM    WAREA,3,1Ø
BP    FOUR2
AM    WAREA,1,1Ø
BP    ONE4
AM    WAREA,1,1Ø
BP    ONE5
AM    WAREA,1,1Ø
BP    ONE1
AM    WAREA,1,1Ø
BP    ONE2
AM    WAREA,1,1Ø
BP    ONE3
AM    WAREA,1,1Ø
BP    FIVEB
AM    WAREA,1,1Ø
BP    FOUR3
AM    WAREA,1,1Ø
BP    ONEB
AM    WAREA,1,1Ø
BP    FIVE1
AM    WAREA,1,1Ø
BP    FIVE2
AM    WAREA,1,1Ø
BP    B4
AM    WAREA,1,1Ø
BP    B5
AM    WAREA,1,1Ø
BP    B1
```

```
11Ø         AM     WAREA,1,1Ø
111         BP     B2
112         AM     WAREA,1,1Ø
113         BP     B3
114         AM     WAREA,1,1Ø
115         BP     3B
116         AM     WAREA,1,1Ø
117         BP     FIVE3
118BB       BTM    CFW,Ø,9
119FOUR4    BTM    CFW,44,9
12ØFOUR5    BTM    CFW,45,9
121TWOB     BTM    CFW,2Ø,9
122FOUR1    BTM    CFW,41,9
123FIVE4    BTM    CFW,54,9
124FIVE5    BTM    CFW,55,9
125THREE4BTM    CFW,34,9
126THREE5BTM    CFW,35,9
127THREE1BTM    CFW,31,9
128THREE2BTM    CFW,32,9
129TWO4     BTM    CFW,24,9
13ØTWO5     BTM    CFW,25,9
131TWO1     BTM    CFW,21,9
132TWO2     BTM    CFW,22,9
133TWO3     BTM    CFW,23,9
134FOURB    BTM    CFW,4Ø,9
135THREE3BTM    CFW,33,9
136FOUR2    BTM    CFW,42,9
137ONE4     BTM    CFW,14,9
138ONE5     BTM    CFW,15,9
139ONE1     BTM    CFW,11,9
14ØONE2     BTM    CFW,12,9
141ONE3     BTM    CFW,13,9
142FIVEB    BTM    CFW,5Ø,9
143FOUR3    BTM    CFW,43,9
144ONEB     BTM    CFW,10,9
145FIVE1    BTM    CFW,51,9
146FIVE2    BTM    CFW,52,9
147B4       BTM    CFW,4,9
148B5       BTM    CFW,5,9
149B1       BTM    CFW,1,9
15ØB2       BTM    CFW,2,9
151B3       BTM    CFW,3,9
1523B       BTM    CFW,3Ø,9
153FIVE3    BTM    CFW,53,9
154NOP      NOP
155CFW      TD     INPUT1+9,NOP+1Ø,2
156TDW      TD     INPUT1+1Ø,NOP+11,2
157         AM     TRANS+11,2,1Ø
158         AM     CFW+6,2,1Ø
159         AM     TDW+6,2,1Ø
16Ø         SM     WOP,1,1Ø
161         BP     TRANS
162         WNCDINPUT1+4
163         BNC1START
164         SF     INPUT1+4
165         TF     INPUT1+83,INPUT1+8
166         CF     INPUT1+79
```

```
167           CF    INPUT1+4
168           WNCDINPUT1+79
169           B     START
17ØINPUT      DAS   8Ø
171           DAC   1,@
172INPUT1     DAS   81
173WOP        DS    3
174WAREA      DS    3
175NB         DNB   3
176           DEND  4Ø2
```

The speed of this program on the 1620 Model I is approximately forty-six cards per minute input when two output cards are punched for each input card. It is much faster when only one output card for one input card is desired. The identification number is always in both cards when two cards are punched for one.

If it is desired to use this program as a subroutine, instructions 6 to 17 inclusive may be eliminated and so forth. Blanks are then translated as zero punches. If an output identification number is not needed, instructions 23 to 40 inclusive and 175 may be eliminated. No 1620 special features hardware is required for this program.

Educational and Psychological Measurement
1966, 26, 739-742.

PUNCHING MULTIRESPONSE QUESTIONS WITH THE IBM 1230 OPTICAL MARK SCORING READER: A PROCEDURE AND AN IBM 1620 SPS COMPUTER PROGRAM

JOHN F. GUGEL
Hunter College

Under ordinary circumstances tests or questionnaires requiring more than one response at the same time to a given question cannot be punched by the IBM 1230 Optical Mark Scoring Reader—534 Key Punch System. The 1230 reads across the answer sheet by looking at ten response positions in a line on the left and by then following the same procedure on the right. When responses are to be punched for a test item analysis or for a questionnaire, the machine does not punch less than the ten positions at one time. If more than one mark occurs in each set of ten positions, a multipunch results. If there are more than three marks to the set, a "12" punch results when the 534 Key Punch is programmed with a "2" punch in the drum card for the column to be punched. The same outcome results if there are more than two when the "3" (alphameric) punch is used on the drum.

However, an answer sheet or questionnaire with more than one possible response per five choice question can be punched. As one faces the side of the IBM 1230 with the operating dials and switches, a panel door on the right side can be seen. When it is opened, a panel of twenty-one numbered buttons with inserts for a screw driver is visible. The buttons are used to change the brightness of the light bulbs that shoot light to the passing answer sheet for reflection back to the photoelectric cells. If the buttons are turned clockwise, the lights get brighter. If turned counterclockwise, the lights get dimmer. (Before turning any of these buttons,

it is a good idea to mark where each button is set so that it can be reset.) If one of these buttons is turned completely counterclockwise, its light goes out and the 1230 ignores all positions over which the bulb passes. For a set of ten bulbs it is possible to turn off nine and keep one on. The punching problem can be solved by making ten passes with each answer sheet, but the identification numbers of the papers would be split up onto several cards.

When identification numbers and faster processing are required, two positions of the sets of ten may be taken at one time with the "3" punch on the 534 drum card used. This calls for five passes at most. Alphameric characters are punched for the two lighted marked positions. (One of the two marks must come from the first group of five question alternatives in the set of ten positions, and the second mark must come from the second group of five—two may not come from the same group of five at the same time.)

The Program

The following 1620 SPS computer program then can be used to put the identification number back together and combine desired combinations back onto one card:

```
001         DORG 402
002ORG      BNI  OFF,900
003OFF      TFM  TR1+11,INPUT1-1
004         TFM  TR2+11,INPUT2-1
005         TFM  TR3+11,INPUT3-1
006         TFM  TR4+11,INPUT4-1
007         TFM  TR5+11,INPUT5-1
008         RNCDHEADER
009         SF   HEADER
010         MM   HEADER+1,2,10
011         SM   99,2,10
012         A    TR1+11,99
013         A    TR2+11,99
014         A    TR3+11,99
015         A    TR4+11,99
016         A    TR5+11,99
017START    TFM  A+6,OUTPUT-1
018         TFM  A+11,INPUT1-1
019         TFM  B+11,OUTPUT-1
020         TFM  C+6,OUTPUT-1
021         TFM  C+11,INPUT2-1
022         TFM  D+11,OUTPUT-1
023         TFM  E+6,OUTPUT-1
024         TFM  E+11,INPUT3-1
025         TFM  F+11,OUTPUT-1
026         TFM  G+6,OUTPUT-1
027         TFM  G+11,INPUT4-1
```

```
028          TFM  H+11,OUTPUT-1
029          TFM  I+6,OUTPUT-1
030          TFM  I+11,INPUT5-1
031          TFM  CUB,9,10
032          RACDINPUT1
033          RACDINPUT2
034          RACDINPUT3
035          RACDINPUT4
036          RACDINPUT5
037TR1       TR   OUTPUT+9,INPUT1-1,7
038TR2       TR   OUTPUT+39,INPUT2-1,7
039TR3       TR   OUTPUT+69,INPUT3-1,7
040TR4       TR   OUTPUT+99,INPUT4-1,7
041TR5       TR   OUTPUT+129,INPUT5-1,7
042A         TD   OUTPUT-1,INPUT1-1,2
043B         BD   MOD,OUTPUT-1,7
044C         TD   OUTPUT-1,INPUT2-1,2
045D         BD   MOD,OUTPUT-1,7
046E         TD   OUTPUT-1,INPUT3-1,2
047F         BD   MOD,OUTPUT-1,7
048G         TD   OUTPUT-1,INPUT4-1,2
049H         BD   MOD,OUTPUT-1,7
050I         TD   OUTPUT-1,INPUT5-1,2
051MOD       AM   A+6,1,10
052          AM   A+11,1,10
053          AM   B+11,1,10
054          AM   C+6,1,10
055          AM   C+11,1,10
056          AM   D+11,1,10
057          AM   E+6,1,10
058          AM   E+11,1,10
059          AM   F+11,1,10
060          AM   G+6,1,10
061          AM   G+11,1,10
062          AM   H+11,1,10
063          AM   I+6,1,10
064          AM   I+11,1,10
065          SM   CUB,1,10
066          BNF  A,CUB
067          WACDOUTPUT
068          BNLCSTART
069          RCTY
070          WATYHDR
071          B    ORG
072HEADERDSS      80
073CUB       DS   2
074OUTPUT DAS     80
075SPACE    DAS   80
076INPUT1  DAS    80
077R1        DAC  1,@
078INPUT2  DAS    80
079R2        DAC  1,@
080INPUT3  DAS    80
081R3        DAC  1,@
082INPUT4  DAS    80
083R4        DAC  1,@
084INPUT5  DAS    80
```

```
085R5      DAC  1,@
086HDR     DAC  17,LOAD NEW HEADER.@
087        DENDORG
```

Columns 1 to 5 of the input and output are reserved for the identification number. The output responses are in groups of fifteen in columns 6 to 80. Input is transferred in groups of 15 columns from each card. The first 15 come from the first card, the next from the second card, and so forth. A header card precedes the data cards. It tells the computer from which column the input transfer is to start. This column is punched in columns 1 and 2 of the header. If, for example, the groups of 15 start in column 6, an "06" is punched in the first two columns. Then columns 6 to 20 of the first data card will go into columns 6 to 20 of the output card; columns 6 to 20 of the second card will go into columns 21 to 35 of the output card; columns 6 to 20 of card 3 will go into columns 36 to 50 of the output card, and so forth. Data input card one is followed by input card two, which is followed by input card three, and so forth. The input cards thus have to be carefully put in order and stacked together by hand.

The author's program for unpacking the special code should now be used. It is recommended that the light pairs set on the 1230 be "1–1", "2–2", "3–3", "4–4" up to "5–5." Then the unpacked cards will have all the 1's from the first cluster of questions followed by all the 2's and so forth.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 743-745.

SCORING MULTIRESPONSE QUESTIONS WITH THE IBM 1230 OPTICAL MARK SCORING READER

JOHN F. GUGEL
Hunter College

UNDER ordinary circumstances tests requiring more than one response at the same time to a given question cannot be scored by the IBM 1230 Optical Mark Scoring Reader. The 1230 reads across the answer sheet by looking at ten response positions in a line on the left and then by doing the same on the right. In the scoring process the machine will score as many as four responses per ten positions if the Question Length 2 setting is used. Response positions three and eight, however, would not be read, and papers with more than one mark per set of two questions would be selected and removed from the other papers.

Nonetheless, questions with several right answers at the same time can be scored if an answer sheet is designed with lettering and numbering going down the page. With one position of each set of two across left blank, the Question Length 2 setting could be used. The author used this procedure in designing the answer sheet in this article from an IBM answer sheet by photo-offset. It was designed for use with the Differential Aptitude Test Battery.

The special answer sheet is scored with the Question Length 2 setting. It must be passed through a second time for a count of wrong answers (if needed). If the full number of questions is used, the number of right answers may be printed on one spot in the right hand margin and the number of wrong answers on the spot next to it. A little peg under the inner top hood of the machine makes it possible to print scores in two different places.

If punching of questions is required as for an item analysis, the procedure should be used that is described in the author's preceding article entitled "Punching Multiresponse Questions With the IBM 1230 Optical Mark Scoring Reader: a Procedure and an IBM 1620 SPS Computer Program."

NAME ______ LAST FIRST MIDDLE ______ GRADE ______ SEX ______ (M OR F) AGE ______ DATE ______

SCHOOL ______ NAME OF TEST ______ INSTRUCTOR ______

J.G.-4

MAKE YOUR MARKS HEAVY AND BLACK— —ERASE COMPLETELY ANY ANSWERS YOU WISH TO CHANGE

1	11	21	31	41	51	61	71
A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5
2	12	22	32	42	52	62	72
A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5
3	13	23	33	43	53	63	73
A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5
4	14	24	34	44	54	64	74
A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5	A 1 B 2 C 3 D 4 E 5
5	15	25	35	45	55	65	75
A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4	A 1 B 2 C 3 D 4

					5 E
6	1 A	2 B	3 C	4 D	5 E
7	1 A	2 B	3 C	4 D	5 E
8	1 A	2 B	3 C	4 D	5 E
9	1 A	2 B	3 C	4 D	5 E
10	1 A	2 B	3 C	4 D	5 E
					5 E
16	1 A	2 B	3 C	4 D	5 E
17	1 A	2 B	3 C	4 D	5 E
18	1 A	2 B	3 C	4 D	5 E
19	1 A	2 B	3 C	4 D	5 E
20	1 A	2 B	3 C	4 D	5 E
					5 E
26	1 A	2 B	3 C	4 D	5 E
27	1 A	2 B	3 C	4 D	5 E
28	1 A	2 B	3 C	4 D	5 E
29	1 A	2 B	3 C	4 D	5 E
30	1 A	2 B	3 C	4 D	5 E
					5 E
36	1 A	2 B	3 C	4 D	5 E
37	1 A	2 B	3 C	4 D	5 E
38	1 A	2 B	3 C	4 D	5 E
39	1 A	2 B	3 C	4 D	5 E
40	1 A	2 B	3 C	4 D	5 E
					5 E
46	1 A	2 B	3 C	4 D	5 E
47	1 A	2 B	3 C	4 D	5 E
48	1 A	2 B	3 C	4 D	5 E
49	1 A	2 B	3 C	4 D	5 E
50	1 A	2 B	3 C	4 D	5 E
					5 E
56	1 A	2 B	3 C	4 D	5 E
57	1 A	2 B	3 C	4 D	5 E
58	1 A	2 B	3 C	4 D	5 E
59	1 A	2 B	3 C	4 D	5 E
60	1 A	2 B	3 C	4 D	5 E
					5 E
66	1 A	2 B	3 C	4 D	5 E
67	1 A	2 B	3 C	4 D	5 E
68	1 A	2 B	3 C	4 D	5 E
69	1 A	2 B	3 C	4 D	5 E
70	1 A	2 B	3 C	4 D	5 E
					5 E
76	1 A	2 B	3 C	4 D	5 E
77	1 A	2 B	3 C	4 D	5 E
78	1 A	2 B	3 C	4 D	5 E
79	1 A	2 B	3 C	4 D	5 E
80	1 A	2 B	3 C	4 D	5 E

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 747-748.

AN ALTERNATE TIME-SAVING PROCEDURE FOR COMPUTING Z SCORES[1]

R. J. RANKIN
Oklahoma State University

HEALTH (1965) reports a fast method for computing z scores suitable for the two keyboard Marchant. The basic z score equation is,

$$z = \frac{X_1 - \bar{X}}{\sigma}(10) + 50,$$

where X_1 is an observed score, $\bar{X}$ is the mean of the observed scores, and σ is the calculated standard deviation of the distribution of observed scores. All short method z score computations are derived from the obvious linear relationship between raw and z scores. Health's method requires computation of a constant $10/\sigma$ and multiplying the raw score difference between a known z score's equivalent raw score and the raw score for which z is desired, then adding the result to the last computed z score to obtain the new z score. This continuous process requires a running account of the distance between adjacent scores. The check requires the computation of the mean of the obtained z scores.

A simpler procedure, suitable for all one keyboard calculators and illustrated with a Monroe model IQ 213 desk calculator, does not require the computation of raw score differences by the operator.

1. Compute the z score for the highest (z_1), and one digit less than the highest raw score (z_2). z_2 may not be equivalent to any

[1] This project was supported by the Research Foundation, Oklahoma State University, Stillwater, Oklahoma.

obtained raw score, but the computed z's must be adjacent in raw score terms.

2. Find the difference between z_1 and z_2, which is a constant (k) for any adjacent scores.
3. Clear the machine and lock the lower dials.
4. At the far left of the keyboard, enter the highest raw score. Skip one column and enter z_1, computed to the desired accuracy.
5. Enter the above from the keyboard to the lower dials.
6. Under the one's place for the raw score, enter 1. At the appropriate decimal place under the z_1 score, enter the constant (k) obtained in Step 2.
7. Make sure the keyboard will not clear automatically by cancelling the "automatic keyboard clear" and lock down the repeat slide into the repeat position.
8. Punch the minus bar. The left position contains the raw score equivalent for z_2, the right position shows z_2.
9. Punch the minus bar to obtain all desired z scores.
10. Check by noting whether the z equivalent for the mean is equal to 50.

With the exception of computing z_1, z_2, and k, the entire process can be completed on the machine with no further computation. It is desirable to arrange the test papers in raw score order, but if one is out of order it is a simple matter to run the + bar or — bar to the desired raw score.

Purists may wish to compute k to 7 places, but this is not necessary unless the range of raw scores is wide. A good check on k is to compute a $z_2 - z_3$ difference which should be equal to $z_1 - z_2$.

Statistics other than z may be computed with this procedure by simply changing the appropriate elements in the formula.

No claim is made for the originality of this method, and the writer is certain that many others faced with the need for computing many z's have equivalent procedures.

REFERENCES

Health, H. A. Time-saving Procedure for Computing Z Scores. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1965, 25, 323–325.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 749-811.

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 751-811.

Standards for Educational and Psychological Tests and Manuals, prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. Washington, D. C.: American Psychological Association, Inc., 1966. Pp. iv + 40. $1.00.

The Professional Service Representative of a Test Publisher

It is the opinion of one test developer and publisher (this one) that *Standards* may well be one of the half dozen most important publications in the history of measurement. Written in clear, concise English that is a tribute to the skill of its several authors, it enunciates fully a set of principles which have been growing for fifty years among the professionals of testing.

This booklet's two predecessor publications—*Technical Recommendations for Psychological Tests and Diagnostic Techniques* published by the American Psychological Association (APA) in 1954 and *Technical Recommendations for Achievement Tests* published by the American Educational Research Association (AERA) and the National Council on Measurements Used in Education in 1955 (now called the National Council on Measurement in Education [NCME])—represented a first and very large step toward codification of professional expectations for published tests. *Standards* is a second and even more impressive stride in the same direction.

The book opens with an excellent description of its own history and purposes and methods, of the audiences who may utilize it, of the cautions to be observed by those who do use it—a modest and wholly objective set of directions for the reader. Then it goes on to present the "standards" in well-organized sections devoted to (a) dissemination of information, (b) interpretation, (c) validity, (d) reliability, (e) administration and scoring, and (f) scales and

[Editor's note: Shortly before the *Standards for Educational and Psychological Tests and Manuals* was released, the editor sought reviews from several individuals with different professional orientations. At the beginning of each review is a caption or heading that describes the editor's categorization of the professional identification of each person who on extremely short notice contributed a review of the *Standards*. Deep appreciation is expressed to each reviewer who participated in evaluating the *Standards*.]

norms. The clarity and comprehensiveness of its coverage of critical topics are sure to gain wide use for this publication in the *teaching* of measurement as well as in the technical assessment of standardized tests.

Not all publishers will agree that all the standards are appropriate, of course. Perhaps not even one publisher will think that every recommendation is both important and well-stated. But nearly all will agree that on every important point the committee has taken a position that is shared by substantial numbers of specialists in the field—a position that can be defended with vigor. This means that the committee actually is speaking for the profession by reflecting its divisions of opinion as well as its more nearly unanimous points of view, instead of legislating its own rules for the publishers. So a listing of the specific points upon which the reviewer holds an opinion differing from that of the committee would amount to little more than an interesting exercise; a different reviewer would produce a different list—and have an equally hard time proving that his position on a given point is better than the committee's.

As one would expect in a publication developed by a committee (even when the final writing is done by one or two very competent people) there is some unevenness in technical sophistication, as well as in expository and editorial styles, from section to section. Compared with the concepts and procedures specified in the section on reliability, for example, the recommendations pertaining to scales and norms are more often unspecific and occasionally over-simplified. And the section on scales and norms would be better with an introduction as useful and well-written as the introductions to the sections on validity and reliability. But this is strictly a *comparative* criticism, for all sections of the book are admirably well done, just as they are.

Continuing the trend which was accelerated by its predecessor publications, this book will make things even more difficult for the individual test-developer as a publisher of operational tests. Authors who act as their own publishers just cannot afford the costs of meeting these standards at the point of publication; by the time they can hope to complete most of the essential standards, the content of their tests will have become obsolete. In this particular effect upon publishing practice, *Standards* offers a fairly accurate reflection of economic and technical circumstances that influence publishing. Development of tests that satisfy current requirements in the field requires large financial investment and enormous technical resources.

Standards should serve its central purpose well. By the time this review appears, surely, the major developers of tests will know the recommendations by heart and will be planning to include in fu-

ture instruments those elements or standards which they have not already built into their publications. Further, measurement technicians who buy and use tests will very quickly build *Standards* into assessment procedures for test materials. Finally, it ought to become evident to authors and teachers in the field that this publication is an admirable statement of proper usage and nomenclature in the language of testing—amounting almost to a manual of style—which should be imitated widely.

In conclusion, any reviewer whose career is related to test publishing should be permitted one small, wistful hope. Partly as a consequence of the efforts of professional groups like APA, AERA, and NCME the technical standards of test publication have improved tremendously in the last two decades. It would be fair to say, however, that as a result of such general improvement a great many tests and manuals now are far better than the uses to which they often are put. A highly sophisticated instrument works no better than a cruder tool in the hands of an apprentice just learning the trade. To put high-standard instruments more often to high-standard uses, then, the technical qualifications of those who choose, administer, and interpret tests should be brought to the technical level of the best tests they employ. Because it is evident that nearly all the recommendations in *Standards* need only the slightest turning or re-phrasing to be wonderfully applicable as *standards for the users of tests,* and because no other group could possibly muster a comparable combination of professional influence and skill, it is to be hoped that the combined committees on test standards will undertake next the development of just such a publication; and make it as good as this one is.

John E. Dobbin
Educational Testing Service

A Measurement Specialist and Executive of a Large Test Publishing Firm

Publication of *Standards for Educational and Psychological Tests and Manuals,* the up-dated version of the *Technical Recommendations* for Psychological Tests and Diagnostic Techniques, and for Achievement Tests, prompts this test publisher to ask himself several questions: What effects have the *Technical Recommendations* had on test publishing in the decade since their appearance? What effects have they had on test selection, interpretation, and use? Are the new *Standards* likely to be better or worse in any of these respects than the original standards?

One approach to the first question would be to compare the manuals for tests published by the major publishers during, say, the past three years with the manuals of tests published in the

three years prior to issuance of the *Recommendations.* While the reviewer has not gone through this exercise, he does not hesitate to assert that the quality of test manuals has, indeed, improved over the past decade, nor to attribute much, perhaps most, of this improvement to publication of the *Standards.* Other influences for good have, to be sure, been operative: the *Mental Measurements Yearbooks,* the general increase in level of sophistication concerning tests, even, paradoxically, the wave of criticism of testing, which has forced the test makers to be on their mettle; but there is no gainsaying the impact of the *Recommendations.*

To the reviewer's knowledge, the *Recommendations* have been very much in the minds of the staffs of the major test publishers, in whose hands lies the responsibility for the preparation of most test manuals. They have perceived the *Recommendations* as a helpful codification of what the profession thinks it reasonable to expect test makers to provide in the way of information about their instruments. "Conscience," someone has written, "is a small voice telling us what we know." The *Standards* may tell the responsible test maker little he does not, or should not know, but it is not a bad thing that he be told, in some organized authoritative way, what his fellow professionals expect of him. Most publishers, the reviewer would judge, view the *Standards* (even those classed as "Essential") as definitions of ideal practice rather than rigid prescriptions. Few, if any, manuals, even among those published in recent years, conform to the letter of the *Standards* in every respect; but it is a fair presumption that where a manual does not accord in every particular with the *Standards,* it is not because the *Standards* were ignored, but rather because a decision was taken, for good reasons or bad, not to abide by the *Recommendations* in the particular instance. This reviewer has sensed no disposition on the part of any major publisher to dismiss the *Standards* as unreasonable, unattainable, or irrelevant; indeed, it is hard to imagine that any serious test author or publisher would lightly disregard a body of recommendations developed with such obvious thought and care, and enjoying the endorsement of the most directly involved professional groups.

It is, of course, not to be expected that every recommendation, or the classification of recommendations as "essential," "very desirable," or "desirable," will command universal endorsement. This reviewer, for example, does not agree totally with the formulations in the revised *Standards* on the use of correction-for-guessing formulas, on provision of tables of equivalence between new and revised forms as a universal practice, on handling the matter of local norms, on treatment of test scores over time, on provision of slope and intercept information in connection with all validity coefficients; he shares the committee's bias in favor of a standard-score

system of reporting and interpreting, but can sympathize with competing views, and so on. Every practitioner will have his own list of places where his judgment differs to some extent from that embodied in the *Standards*. Such lack of total agreement seems to the reviewer unimportant, and perhaps an inevitable consequence of the less-than-fully-developed state of our science. What is important is that the reader sense the goal the committee had in mind in calling for certain kinds of information, and be guided by the spirit of the recommendations.

Beyond the obvious use of the *Standards* as a guide in the preparation of test manuals, this publisher, at least, has found them helpful both with respect to the evaluation of test manuscripts being considered for publication, and in the training of editorial staff.

Less easy to answer are the questions on the effect of the *Recommendations* on test selection, use, and interpretation. There is little evidence to lead the reviewer to believe that many test users are now influenced in their selection of tests by the extent to which their manuals conform to the *Technical Recommendations*. Neither does there appear to be convincing evidence of correlation between improvement in test manuals and improvement in test use. Does this argue against the usefulness of the *Standards*, or call into question the wisdom of the sponsoring bodies in devoting so much time and effort to their production? The reviewer thinks not. He concludes that what is called for is much better dissemination of the *Standards*, and better training of test users in the application of the information which, thanks to the *Standards*, is increasingly available in test manuals. It is unrealistic to pretend that the modal user of educational or psychological tests today is behaving very differently from his counterpart of a decade ago as a result of publication of the *Standards*, and perhaps the *Standards*-writing groups never expected such an outcome. But there are grounds for some small optimism: most measurement textbooks that have appeared in the past five years have devoted attention to the *Standards*, which attention, it is to be hoped, will result in improved test interpretation and use.

How do the revised *Standards* compare with the earlier ones? In structure, organization, scope, and general tone, the revised *Standards* are very similar to the original *Recommendations*—indeed, surprisingly similar in view of the amount of time and study devoted to preparation of the revision. Insofar as the *Standards* constitute a kind of index of the state of the science and the art, one is impressed at how little things seem to have changed in the ten years between the two editions. The treatment of validity, while somewhat expanded and recast, adheres to the concepts of content, construct, predictive and concurrent validity elaborated in the original *Recommendations*, although predictive and con-

current validity have been combined into one category. Most notable changes occur in the section on reliability; here the committee has espoused wholeheartedly an analysis-of-error-variance approach, a development which this reviewer counts as a decided improvement. Recommendations in the areas of administration and scoring, and scales and norms, are much like the original.

Whether or not the revised *Standards* will be more efficacious in improving the quality of test manuals and level of test usage depends less on the changes between the new and the old editions than it does upon what steps are taken to bring the new *Standards* to the attention of test makers and test users. As far as the test makers are concerned, it is safe to assume that the new *Standards* will receive as much attention and observation as did the *Recommendations.* The real task continues to be that of consumer education aimed at increasing the user's ability to utilize the information that the test publishers are being stimulated to provide.

A final word seems in order. To conform fully to the spirit of the *Standards* imposes formidable obligations on the test maker. The development and dissemination of all the information called for is a costly process, and often implies a long period of pre-publication experimentation. If the conscientious publisher is willing to undertake these additional developmental costs, and seeks to recover them, as indeed he must, in the form of higher prices for materials, he would like to feel that he will not thereby be placed at a competitive disadvantage and that the test user, in turn, will recognize an obligation to support better standards of test development by not preferring materials simply because they are less expensive.

Test makers and test users, APA and AERA, are indebted to the committees that labored so hard both on the original *Recommendations* and on the new *Standards.* Their task has been a difficult and perhaps a frustrating one, since in the nature of things the outcomes of their labors are so hard to assess. This reviewer judges that they have discharged their duties with sophistication, technical competence, and good sense. And if the *Standards* now and again seem to set goals that are unrealistic in a world of time and money, well, what is wrong with a dash of idealism?

Roger Lennon
Harcourt, Brace, and World
Test Division

The Professor of Psychology

After comparing the 1966 *Standards* with the 1954 *Recommendations,* the reviewer gained the impression that there are not many major changes. The 1966 Committee worked long and hard. Thus it

is a tribute to the perspicacity of the 1954 group that their work has lasted so well.

The new publication is, however, a little firmer in tone. One facet of this attitude is the change in title from *Recommendations* to *Standards*. Another example is the fact that in the 1954 *Recommendations* there appeared, in a section on Revision and Extension, the statement "The recommendations are intended to be used without reference to any enforcement machinery." No such statement appears in the 1966 *Standards*. Still another instance appears in Item A1.22, which introduces the idea that promotional material for a test should also not be misleading. Likewise, Item A2.31, concerning the necessity for new data when a short form of a test is brought out, has been raised from Very Desirable to Essential.

A new feature is a topical index, and this should be very useful. It appears, too, that it would have been very helpful for the Committee to have prepared one or more outlines for the organization of a test manual, with citations of the paragraphs of the Standards that are relevant to each item. The reviewer had the feeling that standards bearing on a particular portion of a test manual were somewhat scattered. A confirmation of this impression is found in the number of different sections listed under some of the headings in the index.

One prominent modification is in the discussion of validity, where the terms predictive and concurrent validity have been combined and labeled criterion-related validity. The appropriateness of content validity for achievement tests comes in for its proper emphasis, and the discussion of construct validity has been considerably clarified. Apparently twelve years of discussion and controversy about this latter type has borne some fruit.

The section on Reliability has become more sophisticated, with advocacy of analysis of error variance as the most informative approach. Reliability is defined as accuracy, and the reviewer must confess the Committee has not brought him along with them on this point. It seems that a test can be reliable but biased.

One notes with approbation that the audience to whom the *Standards* are directed has been upgraded from competence at the level of a single course in tests and measurements to one of two or three such courses, plus two semesters of statistics. This explicit recognition that not just anyone is capable of judging the merit of a test is a healthy sign.

The reviewer hopes that the shift in these *Standards* toward more rigor will stimulate the testing fraternity also to move toward a more stringent exercise of concern over what appears on the market. There has been reliance upon voluntary cooperation for a long time and, although there has been some progress, sometimes the most ethical of the test producers have been the ones to suffer.

The necessity for some kind of sanctions seems to be indicated both by the current importance of testing in our society and by the criticisms which have grown out of, and fed upon, legitimate public interest in what is going on. Perhaps a first step could be the establishment, by the three bodies issuing these *Standards*, of a "Seal of Approval" for test manuals!

JOHN E. MILHOLLAND
University of Michigan

The Professor of Educational Measurement

This brief set of guidelines for published assessment devices, like its predecessors, is certain to have a marked impact in the testing field. The document is a well integrated and substantially improved revision of two earlier independent, and partially overlapping, efforts by APA and AERA-NCME committees. It is clearly written in language that places minimal demand on the reader's measurement background. The tone of the *Standards* is suggestive, not dictatorial; care is taken neither to dictate practice nor to discourage change and innovation. The principal request throughout the volume is for an honest portrayal of known information about an instrument. The reader will observe that most of the issues raised in the review are of minor consequence, often relating to matters of personal preference.

The format employed closely parallels that of the earlier versions, categorizing the recommendations into six divisions. The initial section on "Dissemination of Information" presents excellent guidelines for promotional literature on published tests. The second section, "Interpretation," makes helpful recommendations regarding information that should be included in test manuals to assist users in making correct interpretations of the test's results, e.g., making the distinction between statistical and practical significance explicit.

The "Validity" section is substantially modified from previous versions. The three primary validity types, content, criterion-related, and construct, are presented in a unified framework in such a way that it is clear to the reader that a complete study of any test typically involves information about all three types. Topic-by-process matrices are recommended for standardized achievement tests. Criterion-related validity is an integration of the former predictive and concurrent categories. The generally excellent presentation of construct validity fails to make one helpful distinction; namely that between the validity of a construct (e.g. test anxiety) versus the adequacy (or validity) of a particular test as a measure of the construct. The recommendation for uniformity in the use of the term "item discrimination" rather than "item validity" for item-total score relationships is to be commended. The reviewer

would like to have seen some suggested guidance regarding the use of particular item discrimination indexes to assist in reducing the undesirable heterogeneity in current practice.

Although the need for cross-validation is clearly indicated, especially when multiple predictors are involved with small samples, the recommendation to apply a correction for shrinkage to multiple correlations when they are to be presented as evidence of criterion-related validity would seem to have been appropriate. The reviewer would have preferred to have seen greater encouragement given to the reporting of confidence intervals and/or standard errors of validity and reliability coefficients. In addition, a suggestion to explore possible non-linear relationships between novel tests or criteria, particularly when concerned with non-cognitive variables would seem to have been in order.

The excellent "Reliability" section emphasizes the determination of components of error variance in tests, a distinct improvement in approach. No longer are coefficients classified into types; the use of suitable descriptive phrases that convey the meaning of reported coefficients is encouraged.

The brief section on "Administration and Scoring" parallels the general high quality of the document. The reviewer would have preferred, instead of recommending that the "correction for guessing" formulas be applied on non-power tests, that experimental data supporting the efficacy of the particular method employed be provided. To the reviewer's knowledge, the relative value of a "correction for guessing" formula in all situations has not been definitively established. It does not seem inconceivable that in some situations the gambling set would be positively correlated with a criterion; hence corrected scores might have less validity than uncorrected scores.

The final section on "Scales and Norms" commendably recommends the use of standard scores in reporting test results more strongly than was done formerly. With respect to the definition and meaning of grade placement scales, the reviewer would have preferred some encouragement toward increased uniformity. Some such direction would probably help *reduce* the unnecessary lack of comparability of grade equivalence from test to test.

In summary, this document is an outstanding example of high quality of content expressed in succinct, non-technical language, both of which will greatly contribute to its usability. Every major objective of the report appears to have been abundantly achieved; in fact, the product might aptly be considered as typifying a standard for such *standards.*

Kenneth D. Hopkins
Laboratory of Educational Research
University of Colorado

The Clinical Psychologist

Clinical psychology, to which the testing movement owes much of its early origins (and present plight), has an important stake in the continuing health and vitality of the movement. Any effort that will help the clinical psychologist to select from the many and varied test procedures now available and in development is bound to have a salutory effect on the currently chaotic scene. The criteria developed for reporting test information in the current *Standards*, if accepted by those who write test manuals, should ease the problem considerably.

Clinicians are singled out for mild rebuke on the problem of quantification of the results of clinical instruments, particularly the projective techniques. The claim that projective techniques cannot be submitted to the same standards employed in evaluating other psychological tests is examined briefly and, in the main, dismissed. Much more on this particular topic can be found in Zubin *et al.* (1965), who argue for the return of projective techniques to the psychometric tradition. An additional implied criticism of test validation of clinical instruments is the collection of data on the basis of availability. The number of studies conducted on the basis of searching through hospital or clinic files after reading an article on a new scoring method for an old technique is beyond calculation.

In addition to these comments, clinical psychologists will find many lively topics in this jam-packed pamphlet which will have particular relevance for testing in clinical situations. The committees, for example, suggest that test manuals, where appropriate, report on the results of fakability studies. This might be expanded to a suggestion that a "faking" key be developed, put through a try-out, and reported. Concern for acquiescence response set or "yes-no" type responses is also expressed, along with suggestions for its detection. The problem of "base rates" for diagnostic instruments appears to have been ignored, although the amount of misclassification or overlapping is listed as an essential item for inclusion in the test manual. How well will a test work in a practical situation if, in the validation study the cut-off score makes very few errors in distinguishing between fifty normals and fifty schizophrenics? While the base rate can be expected to vary from one practical situation to another, the reader of a test manual might be interested in the answer to this question in selected situations.

The problem of validity is simplified into three aspects: content, criterion-related, and construct. Most of the suggestions, as might be expected, are centered on criterion-related validity. Perhaps the one most ignored aspect of the problem is content validity: the nature of the universe being sampled and the adequacy of the sampling. The *Standards* suggest that "the manual should justify the

claim that the test content represents the assumed universe tasks, conditions, or processes (p. 12)." How well does a series of drawings, to which subjects make up stories, represent a universe of situations? What is the universe of items from which questions on a personality questionnaire should be drawn? The reader can see the impossibility of answering these questions at the present time. Psycholinguists are becoming concerned with this problem (Coleman, 1964), and a similar concern may take place in psychological testing.

When test manuals of assessment procedures now employed in clinical situations are evaluated by the criteria in the *Standards*, it will become apparent that clinical psychology will become the chief beneficiary of the new standards. If psychologists of all fields will insist on these standards in the test manuals of the future, the effect cannot help but be healthy and profound.

REFERENCES

Coleman, E. B. Generalizing to a Language Population. *Psychological Reports*, 1964, 14, 219–226.

Zubin, J. Eron, L. D., and Schumer, Florence. *An Experimental Approach to Projective Techniques*. New York: Wiley, 1965.

PHILIP HIMELSTEIN
Texas Western College of the University of Texas

The Measurement and Evaluation Specialist in a Large Public School System

The authors of *Standards for Educational and Psychological Tests and Manuals* are to be commended for their efficient and easy-to-read presentation of so many well conceived recommendations which in some cases utilized some rather involved concepts. Only occasionally are there sections that discuss concepts which exceed the technical competencies of the typical educator. In those areas where educators may encounter cognitive difficulties in measurement concepts the document has presented background information to the reader. In addition, examples and comments are provided to explain the standards. The profession should salute the cooperative efforts of such a distinguished group of mental measurement authorities for their extended dedication to complete the difficult assignment.

To the educator especially, criteria for determining certain aspects of the quality of tests and test manuals are strongly needed because of the variety of backgrounds and training in educational measurement of teachers, principals, and guidance workers and because of the many and sundry tests for school use that have man-

uals with diverse emphases. The document has developed such criteria. Although the publication may be used by the most astute student of mental measurement, it lends itself well to the school teacher, counselor, and administrator. The *Standards* include examples based on subjects ranging from reading readiness tests of the kindergarten child to college and vocational placement tests.

In an era where commercialism has been developed to a high degree in all aspects of business, it is quite appropriate that the publication remind test publishers that manuals, "must avoid using high-pressure advertising techniques." For many of the currently used tests to include all the recommendations, although highly desirable, would pose a rather monumental task on their part. However, it is hoped that test publishers will accept the document as a standard and construct their tests and manuals in keeping with the recommendations in accordance with the standard as conscientiously as they would adhere to a code of ethics.

The format of the document is such that it includes for purposes of clarification through example strengths (as well as deficiencies) of many tests in the field. These models are used frequently as bases for many of the standards.

Although this reviewer feels that the *Standards* could be adhered to without undue stress by test publishers, one statement, (C5.3), has elements of fantasy to expect test publishers to admit that any of their validity samples are, "made up of records accumulated haphazardly." One area untouched by the standard which would be of special assistance to elementary school educators is that of the use of practice tests for students. Primary teachers frequently seek inquiry regarding the use of practice tests prior to the administration of pupils' first tests or types of tests new to pupils.

A concern of the present reviewer is that in section D1.5 from the background of D1.4 a reader may become confused that a statistical significance of difference between two scores of a profile is associated with reliability *per se*. Students of measurement and school staffs who are consumers of tests, all too frequently, become confused from their own self-imposed misconceptions of this difference. Nothing in the *Standards* should enhance this misconception.

The *Standards* not only have presented recommended characteristics which should be included in tests and test manuals but also have, in some cases, indicated challenges or recommended activities that publishers should engage. Two examples which are exemplary of such activities are C7.21 which indicate that new ability tests, "must do more than simply duplicate the measurement of verbal and quantitative ability" and D6.30 which indicates that manuals for general mental ability tests should, "report correlations and

changes in means and standard deviations between tests administered one year apart, two years apart, and three years apart."

The usefulness of the *Standards* to the school workers can be tremendous. The document as a potential to assist educators to assess more effectively tests and manuals is indisputable. However, the previous *Standards* often were not utilized or publicized to the staffs of a large portion of school districts in the country. As these revisions have been made of the 1954 and 1955 *Recommendations,* so too, should ways for implementing the present standards be reviewed. It is the challenge of present educators to develop procedures to publicize and to implement the standards. Educators could publish a supplementary document which would provide additional recommendations for school use. Three examples of materials that could be included in such a document are recommendations for establishing elements of a good testing program, ways to provide in-service training of test interpretation to teachers, and characteristics of a good board-of-education report of a district's test data.

With regard to the myriad of schools and school districts, those agencies which have responsibilities to serve schools have a problem of how to disseminate and implement the standards. Whether the *Standards* become or are used as a tool to improve the use of tests so that schools receive the benefits of the intent of the *Standards* is a function of those staff members in schools and district office personnel who are responsible for testing programs. It is recommended that educators establish local and regional meetings to discuss implications of the *Standards* at their respective level.

Local, regional, and state educational organizations are urged to coordinate the assessment of tests and manuals according to the *Standards.*

Jerry G. Garlock
County Schools of Los Angeles

The Textbook Writer in Measurement and Evaluation

One of the truly helpful documents for test authors, test publishers, and test users was the supplement to the *Psychological Bulletin* entitled *Technical Recommendations for Psychological Tests and Diagnostic Techniques* published in 1954. Beyond a doubt, this publication represents a kind of benchmark in a series of steps leading to the establishment of widely accepted standards for psychological tests and inventories. Today, it is widely mentioned and quoted by the members of the audience to which it was addressed. Moreover, it tended to standardize terminology in testing as it appears in journals, books, and test manuals. Finally, it generated a number of useful manuscripts, including a number of journal ar-

ticles concerning construct validity and a companion publication entitled *Technical Recommendations for Achievement Tests* which appeared in 1955.

Over ten years have passed since the *Technical Recommendations* were published. Now, a revision is available. It, too, is the product of extensive committee work. The product of the efforts of the committee is truly another contribution of significance. The revision concerns both psychological tests and inventories as well as educational tests, and thereby eliminates the need for the two existing publications.

The audience for the revision is described as one which is fairly sophisticated in testing matters. The committee intended that the material included be meaningful to those who have a level of formal training between the master's degree and doctorate in education or psychology at a superior university. Obviously, the classroom teacher who may have had little or no formal education in testing is excluded. This is regrettable, since this very large group is deeply involved in educational testing today, and should have an adequate grasp of major points included in the publication. Nevertheless, the elimination of this audience by the committee is understandable. Now, it is the task of textbooks in educational and psychological testing to capture the ideas prepared by the committee and present them to untrained teachers in such a way that they will be competent users of tests.

The principal section of the revision is the section concerning the standards themselves. As in the case of the 1954 publication, six major subdivisions are included; namely, Dissemination of Information, Interpretation, Validity, Reliability, Administration and Scoring, and Scales and Norms. The most significant of the six are the two dealing with validity and reliability. The remaining are shorter and in some ways less important.

The subdivision concerning dissemination of information and the subdivision concerning interpretation resemble greatly those of the 1954 publication. Two observations are in order, however. First of all, the point is made in several instances that test manuals should indicate clearly that which is *not* measured by the scores yielded. Warning should be given with regard to that which might be thought of as being measured by a test score, but which actually is not. Realistic cautions of this type to the casual user of tests are invaluable. It would be refreshing, indeed, if test manuals would point out needed evidence concerning the usefulness of the test which is missing.

Secondly, one cannot help but note Section A-2 which states that "it would appear proper in most circumstances for the publisher to withdraw a test from the market if the manual is 15 or more years old and no revision can be obtained." Although one can

quarrel with the limitations of 15 years, one cannot quarrel with the intent of this standard. Revisions of tests and test manuals often appear too infrequently.

Even the casual reader of the *Standards* will quickly discover that the discussion of validity is changed in one major way. Whereas the 1954 publication cited four kinds of validity, the revised publication mentions three. Content and construct validity remain virtually unchanged. The original concurrent and predictive validity had been combined into one category known as criterion-related validity. This makes sense, since the only important difference between concurrent and predictive validity is the time element. Emphasis on criterion-related validity is consistent with certain recent publications concerning criterion-related research. Such a categorization of research and categorization of validity reduces the need for detailed description of both.

A point of major emphasis in the validity standards as well as the reliability standards is sampling. Certainly, there has been a tendency by test authors and test publishers to underestimate the importance of this feature of their tests. Samples of overt behavior, samples of subject-matter areas, and samples of subjects are sometimes mentioned briefly with little attempt to defend the strength of the sampling process used. Indeed, sometimes the population which the sample is thought to represent is barely described. The significance of this point is presented well by several standards concerning criterion-related validity. Important criteria to be applied in evaluating the worth of criterion-related validity determinations are listed.

As in the case of the validity subdivision, a noteworthy change has taken place in the reliability subdivision. In addition to modernizing parts of the discussion, the committee decided to abandon some of the terminology used in the 1954 report. Rather than speaking of coefficients of equivalence, coefficients of stability, and the like, it prefers a more complete statement about the reliability coefficient. This statement would briefly describe the nature of the reliability determination, and would not necessarily give it a title such as those previously used.

The section concerning administration and scoring is one of the shorter sections and is essentially complete insofar as the audience of the manuscript is concerned. This brevity, however, should not be interpreted to mean that the standards listed are necessarily of less importance than others. Some of the practices found in schools with regard to test administration and scoring are deplorable. Statements in test manuals are often clear and emphatic. Somehow, however, the message is partially or totally lost to some teachers. One of the frustrations of the testing profession is its inability to communicate with the test users—for example, teachers—with re-

gard to the importance of standardizing administration and scoring practices.

In short, it must be said that the 1966 revision is a fine improvement over the 1954 publication. One of the strengths of the revision yet unmentioned is the set of parenthetical comments which provide examples of violations of a particular standard or successful adherence to it. Although the 1954 publication used this technique as well, the parenthetical comments are more numerous and more pertinent in the 1966 revision.

The addition of the index is excellent. Although it is not an essential part of the publication, it certainly is a helpful part. Another useful addition might have been a selected bibliography concerning standards for tests and manuals. It is obvious that particularly good readings exist in the areas such as validity and reliability. Perhaps in another decade still another committee will examine the current edition and see fit to make this improvement as well.

J. Stanley Ahmann
Colorado State University

Personnel Psychologist in a Large Industry

Possibly never in the history of psychology has so much cogent content been packed in so few pages as in *Standards for Educational and Psychological Tests and Manuals*. In this document, every sentence is highly meaningful, and every example or comment appears necessary for clarity. The authors are to be commended for their conciseness without loss of meaningfulness. The publication represents a considerable improvement over its predecessor, *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, although this earlier document was of extreme value in clarifying many of the issues involving psychological tests.

The strongest features of the new publication lie in the discussions of and the standards for validity and reliability. The discussion on each of these topics should be read by every person who uses tests in any way. The material in these sections has relevance for any psychological research involving tests and certainly should not be, in applicability, limited to test publishers.

It is noteworthy that the interrelations of various types of validity are stressed and that it has been suggested that an earlier classificatory system for reliability coefficients be abandoned. If test publishers follow the new recommendations for reporting reliability data, some of the problems the test user has often previously faced in interpreting such data will become minimal. It is also notable that the section on construct validity is considerably

augmented in the new publication; the cautions mentioned relative to construct validity should be carefully considered by every test publisher and user.

Some persons actively engaged in employment psychology may not feel that *Standards for Educational and Psychological Tests and Manuals* is responsive enough to some of the social and technical issues facing the personnel psychologist today. It is to be noted, however, that it is specifically stated that, ". . . primary responsibility for improvement of testing rests on the shoulders of test users." A further consideration is that some of the presently most pressing problems in employment psychology are likely to be of a temporary nature and will probably not be so important when educational opportunity is more nearly equalized throughout all socioeconomic levels.

There are, however, in the document several points which every personnel psychologist should consider. Among these are strictures against implying that a test measures an "innate" ability and using the labels, "culture-free" and "culture-fair." The cautions expressed relative to the use of moderator variables should be exercised in employment research.

From an editorial viewpoint, the publication is excellent. A rather complete index will aid in making *Standards for Educational and Psychological Tests* a more useful reference than its predecessor.

MARY L. TENOPYR
North American Aviation, Inc.

Language Testing: The Construction and Use of Foreign Language Tests by Robert Lado. New York: McGraw-Hill Book Co., 1964. Pp. xxiii + 389.

In an age of such unprecedented and universal change as ours it may seem presumptuous and pointless to single out change in any one educational field. Still, the upsurge of interest in language teaching and the curricular changes resulting from it present one aspect that is probably unique. While few if any American educators ever seriously advocated dropping science, or mathematics, or social studies, from the secondary school curriculum, a sizable part of the educational community considered the teaching of foreign languages, classical or modern, largely a waste of time. The events of the last thirty years have taught us that the world outside our borders cannot safely be ignored and that, in order to coexist with other peoples, it is necessary to understand and speak their languages. When we took a good look at the type of language teaching going on in our schools, however, we were dismayed to discover that it was indeed a waste of time. The student who left college—

not to mention high school—with a functional knowledge of a language he had studied was the exception rather than the rule.

The effectiveness of recent curriculum changes is due to the fact that they are not superficial reforms of traditional methods but are rooted in a basic new understanding of the nature of language. Modern linguistic science has struck down the artificial categories into which language had been pressed many centuries ago and created new concepts and categories to fit language as it is. The new methods arising from this new understanding are increasingly applied in schools throughout the country. It is the purpose of Dr. Lado's book to provide teachers with adequate techniques for evaluating student learning in courses utilizing these new methods. A well-known linguist, Dean of the Institute of Languages and Linguistics at Georgetown University, and the author of *Language Teaching: A Scientific Approach*, Dr. Lado is eminently fitted for this task. His book is very well organized. Every aspect of the field is covered fully and concisely from the point of view of linguistics, teaching methods, and measurement. A detailed content outline covering 15 pages will effectively guide the teacher seeking information on any specific problem. Part I provides an orientation to the linguistic thinking as well as to the premises about language testing on which the book is based. The teacher unfamiliar with linguistics might do well to peruse a few of the reference works listed before attempting to put the book to practical use. Parts II and III discuss in detail the variables to be tested as well as suitable techniques for testing them. Part IV deals with the special problems of testing cross-cultural understanding which, while transcending language, is yet intimately related to it. The technical aspects of measurement are explained in detail in Part V.

The author points out that some of the techniques generally used in language testing, such as translation, dictation, or essay writing, have little relevance to the skills they claim to measure, and that the same is true for objective tests constructed without a systematic understanding of the variables involved in language learning. Linguistic analysis has made available the means for accurately identifying the elements of language as such as well as those elements integrated in the recognition and production skills. Teachers can now isolate the problems of language learning and test them with a precision far beyond that achieved by the haphazard techniques hallowed by tradition. In their place, teachers will find in this book a variety of techniques to fit different needs, and an alert teacher will be stimulated to find additional techniques of his own. The author takes little for granted. His directions for constructing tests as well as for scoring, norming, item analysis, and other measurement procedures are clear and precise. Advantages and limitations of techniques are pointed out as well as some of the

pitfalls of item writing; some others, however, have been overlooked.

The book has other weaknesses. It is an unchanged reprint of the London edition of 1961 and, judging from the date of the author's preface (1959), it is even older than that. In a field that has advanced rapidly in the past few years this is a serious shortcoming. The two major instruments of standardized foreign language testing, the Modern Language Association's Proficiency and Cooperative tests, which incorporate many of the ideas presented by Dr. Lado and are by now widely used, are not mentioned. Nor are the many tests included in recent textbooks. Surely a few short paragraphs on developments as crucial as these could, and should, have been added. Other references are similarly out of date. The value of the book would have been greatly enhanced by incorporating in it the results of the nationwide experience with standardized, normed test batteries, for instance in the area of constructing and scoring Speaking tests.

Much of the experience on which the book is based was gained through the teaching of English as a foreign language. Teachers of foreign languages should not be deterred by this. It may make them aware of problems in their own teaching which they had not been conscious of before. The few instances in which it has led to inappropriate conclusions can be disregarded. However, teachers may be seriously discouraged by the insistent recommendation to validate tests with native speakers. This is unrealistic except where a language is being taught in its native environment. It also has a number of pitfalls and limitations that are only incompletely pointed out, while most of its benefits can be procured by having a test reviewed by one or two qualified language teachers. Also, teachers familiar with current theories of modern language learning will be taken aback by the frequent use of the native language as a stimulus.

The book has no index. Although it is largely replaced by the content outline, the teacher seeking information on some topic treated earlier would profit from page references. This drawback of the book's praiseworthy conciseness could easily have been avoided.

Not all of Dr. Lado's suggestions can be accepted without qualifications. Factors extraneous to the problem tested are pointed out in some techniques but not in others, e.g. lexical recognition in testing production by sentence completion or the omitted letters technique. While the professional test writer might be able to suggest more appropriate techniques in some cases, or to point out some flaws in item writing that might have been avoided, even he can profit greatly from the theoretical and practical ideas presented in this book.

The author makes it clear that this is a teacher's book, and it

will indeed be invaluable to the language teacher. For the language supervisor, it will be of great help in the construction and use of more formal tests for which the classroom teacher lacks both time and facilities. It should also be recommended reading for administrators and guidance experts. The essential difference between language as a vehicle for the communication of ideas and other subjects that consist of ideas as such, the fundamental differences between testing of the native or of a foreign tongue and, above all, the realization of the many and complex variables that enter into language learning, should come as an eye opener to those who have thought of a foreign language as just another subject, and of a foreign language test as just another achievement test.

SUZANNE STAHL
Educational Testing Service

Acquiescence and Social Desirability Response Styles, Item Characteristics, and Conformity by Lawrence J. Stricker. Psychological Reports Monograph Supplement 2-V12, Missoula, Montana: Southern Universities Press, 1963. Pp. 22. $1.00.

The research reported attempts to integrate conceptualizations from two quite different origins: experimental studies of conformity and personality research on the response styles of acquiescence and social desirability. The first of these concepts involves a subject's yielding to opinions expressed by others in the experimental situation and the second involves the subject's either generally agreeing with paper-and-pencil test items or agreeing with those items that reflect socially acceptable norms. Stricker argues that pre-established norms which a subject carries around with him may affect his responses on personality and attitude tests in the ways that the norms established by experimental "stooges" affect responses in conformity experiments. If so, then there also may be an equivalence between parameters varied in these two types of research. And this is Stricker's most interesting notion.

Ambiguity of objective-structure in the conformity experiment may operate in parallel fashion to yield low-readability of test items. Increasing the size of majority effects in the conformity experiment may have parallel effects increasing the extremeness of socially acceptable personality and attitude items. In addition, Stricker wishes to explore the generality of conforming as a general personality characteristic. If there are those who consistently conform, they should show it not only by agreeing or disagreeing with items according to their social acceptability, but also by simply agreeing with any item for which there is no identifiable social norm.

Subjects were tested in order to classify them on degree of acquiescence and on degree of social desirability response styles. In

all, 27 male and 66 female graduate and undergraduate psychology students were used. Forty-six subjects were dropped because of some "awareness" of the experiment's purpose (a procedure that might have produced some sort of harmful subject selection). At the same time, subjects were answering the experimental items, which consisted of 73 derived from the Social Adjustment subtest of the Minnesota Multiphase Personality Inventory (MMPI) and 65 derived from Ferguson's Humanitarianism scale. There were 16 variations in the items based on four variables: clearness, extremeness, social desirability, and positive vs. negative phrasing. Independent manipulation of the last two variables allowed Stricker to untangle acquiescent responses from social conformity responses.

In the main, Stricker found powerful effects of item variation and virtually no effect of personality-response style. More acquiescent responses and fewer socially desirable responses were made to *unclear* than to *clear attitude* items. This outcome would be expected, since unclear items do not allow the subject to identify social norms to which he might conform. With the *personality* items, fewer socially conforming responses were made to unclear items than to clear items; fewer acquiescent responses were also made to these items. This result fails to agree with the hypothesis. More acquiescent responses were made to moderate personality and attitude items than to extreme ones, again presumably because it is more difficult to identify social norms with moderate items. However, although more socially conforming responses were made to extreme personality items than to moderate ones, no reliable differences were found for extreme and moderate *attitude* items. This outcome, again, is contrary to hypothesis.

Stricker had hypothesized that there should be a strong positive correlation between the two kinds of responses that show most conformity: acquiescent responses to moderate items and socially conforming responses to extreme items. The obtained correlation of —.21 was significant at the .05 level. Taken together with absence of reliable differences as a function of either of the personality response-styles, this result casts doubt on the conception of a general personality dimension of conforming. Perhaps, after all, conforming like suggestability is more a function of the conditions of decision than of the individual's long-term characteristics. Stricker cites many previous studies, the findings on which show that these two response-styles do not relate to amount of conformity in experimental manipulations. Thus the situational point of view is supported.

Although Stricker discusses theoretical possibilities regarding the inconsistencies between his hypotheses and data, it seems likely that experimental artifacts may have produced them. For one

thing, his method of obtaining moderate and extreme items and his criteria for assessing them are not very valid. Many times items were made "more extreme" by adding extreme frequency words like "always" and imperatives like "must." Such a procedure does not really produce a more extreme social attitude; it is the same old attitude held more extremely. The items were then classified by 11 graduate students in psychology for extremeness, clearness, and social desirability—and again for social desirability by 40 undergraduates. Samples of such size are not really large enough for one to formulate important judgments. Indeed, Stricker found: (a) that the two samples disagreed radically on the social desirability of the attitude items, (b) that the clear and unclear items did not differ—either on the Flesch or on the Dale-Chall readability index, and (c) that (most damaging of all) the socially desirable items were rated more "moderate" than the socially undesirable items and vice versa.

Therefore, one might expect: (1) The measurement of a number of socially desirable responses was not completely reliable for the experimental subjects. (2) The effect of item clearness was rather vitiated in the experiment. (3) Since moderateness per se of the items in this study was more socially desirable than was an extreme position, the subject should be more likely to agree than to disagree with any moderate item (artifactually). Moreover, there should be an artificial depression in the number of socially conforming responses to moderate items (since "undesirable, moderate" items were not really so undesirable as might be anticipated).

ALBERT SILVERSTEIN
University of Rhode Island

Psychodiagnostic Testing: An Empirical Approach by Molly Harrower. Springfield, Illinois: Charles C. Thomas, 1965. Pp. xxi + 90. $4.75.

Dr. Harrower had an intriguing idea, one which must have occurred to every psychologist engaged in the practice of individual psychological testing: "What has happened to all the patients I have tested down through the years?" While this question may have caused a momentary stir for most psychologists, Dr. Harrower set out to find the answer. The result is this slim volume, a follow-up study of tested patients referred to Dr. Harrower by psychotherapists of various persuasions.

The purpose is certainly heroic. By analyzing the test records after at least four years has elapsed since testing was accomplished, the author is attempting to determine the test predictors of the

outcome of psychotherapy. Dr. Harrower was able to put together a sample of 622 subjects, an awesome size for a study with this purpose. From the records the author devised a rating scale, the details of which make replication almost impossible.

The test battery consisted of the Wechsler-Bellevue, Rorschach, Draw-A-Person, Sentence Completion, Thematic Apperception Test, The Most Unpleasant Concept, and the Szondi. Just how these varied procedures, some with minimum validity and others considered to be monuments to the *naivité* of clinical psychologists in the area of psychometrics, play a part in what comes later is hardly mentioned. Based in some way, but not detailed, on test performance, the subjects were divided into degrees of "mental health potential." Dr. Harrower attempts to divide mental health potential into categories similar to those used in categorizing subjects tested with intelligence tests (*ie.*, "average," "dull," "bright normal," etc.). This effort to overcome the problem of sorting the results of projective test into psychiatric categories, however, overlooks one important problem. Wechsler's scheme for divisions of intellectual status, the basis for Harrower's hope, is firmly tied to the properties of the normal curve. Does this projective typology share the same property?

Although no cross-validation of the results (subjects with positive mental health are rated as more successful in psychotherapy by the therapists) is presented, the author did check the concept in a different problem, the selection of ministers. Unfortunately, no correlations or significant levels are reported, although the "eyeball" evidence appears to be in line with her notion that mental health potential is an important construct.

Of interest to clinical psychologists is the material in this book devoted to the results of psychotherapy. In general, the trend is marked for orthodox psychoanalysts to rate themselves as most successful, and with departures from orthodoxy, the claims for therapeutic success are reduced. This discussion of therapeutic success is difficult to evaluate. Out of Eysenck's disparaging survey of the results of psychotherapy has come an appreciation of the methodological difficulties in evaluating therapeutic success. Unfortunately for the reviewer, Lawrence S. Kubie's foreword disarmingly admits to many defects in Harrower's study. A few others might also be thrown in: lack of control of the experience level of the therapists, a *laissez faire* approach to the problem of criteria of therapeutic effectiveness, and an avoidance of the issue of control groups. To be sure, it would be difficult to devise a meaningful control group (such as "no treatment") for a study based on patients referred by psychotherapists. But surely the author's files contain many subjects not referred for pre-psychotherapy evaluation and who received a similar test battery. Some method for follow-up

could be developed, with both groups being rated by someone else besides the therapist for the experimental group.

It is the reviewer's opinion that this overpriced book might better have been presented in the form of a journal article. With none of the page restrictions faced by a journal contributor, the insufficient description of the procedures, along with sketchy case-history material on validation, does not justify the hard cover.

PHILIP HIMELSTEIN
Texas Western College of the University of Texas

Basic Statistical Methods (Second Edition) by N. M. Downie and R. W. Heath. New York: Harper & Row, 1965. Pp. xv + 325. $6.95.

Basic Statistical Methods is intended as an introductory text for social science students who generally have a limited mathematical orientation. This objective is successfully attained, since the treatment is clearly written and quite comprehensible. The chapters are short and devoid of verbiage; a real boon. As the print contrasts well with the dull-surfaced paper, eye strain is minimized. Appropriate, relevant, and easy-to-read tables are furnished in the Appendix. Its format is such that a fairly good balance of attention is assigned to describing assumptions, presenting sufficient but not tedious derivations for adequate comprehension of formulae and relationships, citing common formulas with illustrative computations, furnishing usual probability interpretations, and including some practical applications.

Favorable reactions to the text were obtained from an education class of sixteen teachers and/or counselors. Their major comments reflected their pleasure with the light treatment of derivation, with the generally easy-to-follow meaningful examples, and with the readability. Some students thought that smaller numbers could be used in certain assigned problems so that more study time could be applied to understanding and application rather than to concentrating on the calculations. Moreover, they expressed a desire to have the numbers or letters labelling certain problems, conditions, or groups relate to real life situations and not to function only as symbolic designations.

The workbook was unavailable for use by the current class until midway through the semester; consequently it was not utilized. However, it was noted that problem solutions are missing from the workbook. In addition, an instructor's handbook apparently has not been published; at any rate one was not supplied when requested.

Although this reviewer is well satisfied with the text and will be

likely to continue using it with future classes, some suggestions and constructive comments are offered as a result of his experiences with the text this past semester. In general, more practical applications from the social sciences should be included. Any designation of a group or condition should be implemented with descriptive, real life examples, not merely by use of the letters A, B, and C or of 1, 2, and 3 type labels.

The eighteen chapters comprise the main topics usually found in most introductory statistical texts except for the chapter on nonparametric statistics. (A brief introduction, which lists the reasons for studying statistics, is followed by a short history of contributors to statistics. A section on the usefulness of statistics rounds out Chapter 1.)

In Chapter 2 the typical arithmetic fundamentals, types of measurements, measurement scales, and statistical symbols are treated. Only one topic caused minor concern: the term significant digits is never defined although practical, obvious examples for depicting the rules are shown; thus the definition may be inferred. In terms of reducing the number of statistical symbols to differentiate sample or population values, one might simply use a tilde over the symbol for signifying parameter values and thereby reduce the total number of symbols by one half.

In Chapter 5 on variability, a label to identify the first row of numbers in Figures 5.1 and 5.2 as raw scores would help. In Figure 5.3 the order of the scores from high to low could be inverted advantageously to low to high to coincide with the graphic description on "error of coarse grouping" as is ordinarily the case. Rather than imply, the authors should explicitly state that error of coarse grouping tends to enlarge the estimates of sigma and that Sheppard's correction provides the corrected value. An example would be helpful.

The section on "Correlation" (Chapter 7) could have a segment specifically devoted to an interpretation of the meaning of this term. The Pearson r is thoroughly described in Chapter 7, but its interpretation and significance are not spelled out until Chapter 13. With reference to the table of significant values of r, discussion that the probability of an observed correlation as large as or larger than the one found being due to chance, would be suitable implementation.

Chapter 8, "Linear Regression," erroneously calls b_{yx} a Beta coefficient (p. 98). Two additional formulas are suggested for inclusion to complete the methods of computing the b-coefficients because of their common usage: $b_{yx} = r_{yx}\, s_y/s_x$ and its counterpart for b_{xy}.

The formula, $s_f = \sqrt{Npq}$, from Chapter 10, "Sampling," should be related to the section, "Use of the Formulas of the Binomial

Distribution," of Chapter 9, Probability and the Binomial Distribution. Practical examples dealing with smoking preferences, coca cola tasting, and number of items right in objective tests are suitably depicted with the binomial formula ($\sigma = \sqrt{npq}$) in Chapter 9, but its counterpart is glossed over in Chapter 10.

Particularly well-written, Chapter 11, "Testing Hypotheses, The Differences between Means," might be improved in one minor way. It could be specified that t or z tests apply (depending on sample size) to tests of significance of differences between means for correlated data (computed directly from differences without computing a correlation coefficient) in the same section in which the means and correlation coefficient are given rather than two pages later. A table with two columns listing "Conditions" in one column (small or large samples, correlated or uncorrelated data, etc.) plus the "Formulas" in a second column could succinctly summarize the entire chapter. In fact, such a table was constructed as a series of matching items in part of one of the reviewer's course examination. In the latter table, however, actual sample sizes, methods of sample selection, and other information were substituted for the "conditions" suggested above in order to promote understanding, not mere recapitulation of facts.

Perhaps a better appreciation of the Lawshe-Baker Nomograph for testing the significance of differences between percentages or proportions could be accomplished by transferring it from Chapter 12 to Chapter 17, the chapter specifically dealing with the topic of item analysis. Another welcome addition, in Chapter 12 might be McNemar's $z = (f_1 - f_2)/\sqrt{f_1 + f_2}$ as a z-test of departure of two frequencies from equality of frequencies in the diagonal reflecting changes in responses.

An exceptionally clear presentation is made in the "analysis of variance" portion of Chapter 15. Unless a more complete description of simple covariance including an example is presented, the topic might properly have been omitted without serious loss.

Spearman's rank-order correlation coefficient Rho and Kendall's coefficient of concordance W, which are contained in Chapter 16, "Other Correlational Techniques" might more advantageously be placed after the chapter on correlation for the following reasons. Most research studies carried out by teachers or counselors involve sample sizes rarely exceeding 30 or 35 cases. In addition, the convenience and ease of computation of these two statistics to establish evidences of relationships and/or reliabilities of behavior rankings are especially pertinent. Finally, by using the same sets of numbers, where possible, the relative advantages of the three statistics could be more lucidly demonstrated.

Most of the suggestions noted are proposed with the main idea that many social science students enroll in only one statistics

course. Their primary interests are with applications involving data about people, not with abstract research concepts or numbers. They have generally avoided courses emphasizing research methodology and statistical thinking in the past. If statistics is featured in less foreboding ways, probably some of their resistance would dissipate and more research application might result. Many students who have excellent ideas for investigation are reluctant to pursue them because of the numerical "bugaboo." It is probably more efficacious to make the subject of statistics as palatable as possible and still to meet many of their practical needs. Another important reason for a nonthreatening presentation is that students can become improved consumers of published works—consumers who are more aware of cogent questions to be raised when evaluating published research. This point is stressed, since many of these same students eventually become administrators who effect decisions of varying and costly magnitude even though they lack even minimal statistical knowledge. In its present form this text should make definite contributions toward realization of these desired teaching objectives.

Robert R. Morman
California State College at Los Angeles

Introduction to the Theory of Statistics by Alexander M. Mood and Franklin A. Graybill. New York: McGraw-Hill Book Company, 1963. Pp. xv + 443. $8.95.

This book is the second edition of a very popular introductory mathematical statistics book. The first edition developed from the lecture notes prepared for a course given to seniors and first year graduate students by the first author while he was at Iowa State University. Neither edition assumes any statistical background, but each does assume one year of undergraduate calculus. However, the merits of the books, especially this newer edition, will be better appreciated by those readers who have had a previous course in introductory statistics. Matrix algebra is used quite heavily in discussing the multivariate distribution, multiple regression, and experimental design; and although a section is devoted to the elementary properties of matrices and determinants, the reader would probably benefit by having a little prior knowledge of matrices and determinants.

The text would probably be of little interest to the average behavioral or social science graduate student because of the relatively rigorous mathematical treatment of the subject matter. On the other hand, the graduate student who is anticipating an experimentally oriented research career, or the student and prac-

titioner of psychological statistics would certainly do well to read this volume. This introductory book fills many gaping holes that appear in the average statistical texts oriented toward the behavioral and social sciences. Many of the statistical concepts found in this book should also be found in the psychological statistics texts, but disappointingly this is rarely the case.

The volume is comprised of sixteen chapters. Each chapter contains numerous illustrations, ample problems, and a welcome up-to-date bibliography. In the review of this book an attempt will be made to emphasize those topics that rarely appear or are given inadequate coverage in most of the psychological statistics texts, but are given lucid and comprehensive treatment in Mood and Graybill's text. The chapter on probability, which is fairly comprehensive, covers such topics as set theory, sample spaces, marginal probabilities, conditional probabilities, and other fundamental concepts of probability. Since probability is the cornerstone of statistics, a fairly comprehensive discussion of this subject is in order. There is a chapter for both discrete and continuous random distributions. The multivariate distribution is discussed for both cases as well as for marginal distributions. The concept of the moment generating function which is an important tool for defining distributions is discussed and applied continuously throughout the book.

The chapter on point estimation is one of the most unique and informative chapters in the text. In this chapter, the authors discuss such fundamentally important concepts as sufficient statistics, and minimum-variance unbiased estimates. The sections on likelihood estimation and Bayesian estimation are extremely well written. One whole chapter is devoted to the multivariate normal distribution with a subsection on quadratic forms. This chapter is discussed realistically in terms of matrices and determinants, as it should be. With the increasing emphasis on multivariate techniques, this treatment becomes a valuable part of the book. The chapter on sampling distributions not only gives a clear picture of how the distributions of various estimators are derived, but also shows the development of some of the test distributions such as Chi Square and F. The chapter on hypothesis testing is a major contribution in that hypothesis testing is discussed in the framework of modern decision theory rather than in the older classical tradition. Likelihood ratio and goodness of fit tests are also discussed in this chapter.

It is refreshing to see multiple regression and experimental design developed in the framework of matrix algebra. However, it is rather disappointing that experimental design was given such short coverage.

In summary, the reviewer recommends the book to the serious student of psychological statistics. The text gives a good basic

theoretical background of the development of statistics that will lead to more rigorous applications of psychological statistics as well as a deeper appreciation for the discipline of statistics. The breadth, depth of coverage, and currency of the book also add to its attractiveness.

George H. Dunteman
Regional Rehabilitation Research Institute
University of Florida

Linear Algebra by George Hadley. Reading, Massachusetts: Addison-Wesley Publishing Company, 1961. Pp. ix + 290. $7.95.

It was Hadley's intention of fullfilling the students' needs for a knowledge of linear algebra in a number of fields such as engineering and operations research as well as the social and behavioral sciences. Although the book is aimed at students with a limited mathematical background, the average graduate student in the social and behavioral sciences has probably not obtained this "limited" mathematical background. With the increased emphasis on multivariate procedures such as factor analysis, discriminant analysis, and multivariate analysis of variance, some knowledge of matrix algebra, the theory of determinants, and n-dimensional geometry is becoming more essential for research workers and psychological statisticians. This book does an excellent job of presenting some of these basic concepts so essential for an understanding of multivariate analyses. Although other books such as Thurstone's (1947) *Multiple Factor Analysis* and Harmon's (1960) *Modern Factor Analysis* were written primarily for a psychological audience and contain introductory chapters on n-dimensional geometry and the theory of matrices and determinants, their coverage of this subject matter is neither comprehensive nor rigorous enough for the purposes of most students of multivariate analysis. On the other hand, treatments of this subject matter by Anderson (1958) and Rao (1965) seem to be at a too rigorous and abstract level for most researchers. A skillful writer, Hadley presents his topics lucidly at an intermediate level. There are problems and suggested readings at the end of each chapter.

The book is comprised of seven chapters. Chapter 1, the introduction, illustrates the use of linear algebra in the building of linear models. Chapter 2 is concerned with vectors and covers many concepts fundamental to factor analysis and profile analysis such as the magnitude and direction of vectors in n-space, the Euclidean coordinate system, the concept of distance in n-space, and the linear dependence of vectors.

Chapter 3 covers matrices and determinants. In this chapter most of the basic matrices and matrix operations are discussed and

illustrated. The section of the chapter concerned with determinants is especially informative. Two ways to evaluate determinants are discussed while the author is focusing upon the theory of determinants and at the same time providing numerical illustrations. Matrix inversion also receives comprehensive treatment. Many of the matrix operations described in this chapter are used in factor analysis.

Linear transformations and the rank of matrices are discussed in Chapter 4. Chapter 5 is involved with the theory and solutions of simultaneous linear equations. This chapter is probably of less interest and relevance than any of the other chapters. Chapter 6 is concerned with convex sets and n-dimensional geometry and is probably the most valuable and unique chapter, since the reviewer knows of no other books that satisfactorily discuss n-dimensional geometry at a relatively concrete and elementary level. The discussion of set theory will be of special interest to researchers concerned with decision theory and discriminant analysis classification procedures, since essential concepts such as boundaries between regions and hyperplanes are discussed.

The last chapter, Chapter 7, covers characteristic value problems and quadratic forms. The basic determinantal equation that underlies principal component analysis is fully discussed. The numerical illustration of determining latent roots and vectors is especially illuminating. The discussion of quadratic forms should be helpful to the researcher concerned with discriminant analysis and multivariate frequency functions.

In summary, Hadley's book is an extremely well written and comprehensive reference book on the basic formulations of matrices, determinants, and n-dimensional geometry that should prove especially useful for students and practitioners of multivariate analysis. The book in many respects seems almost tailor-made for the psychological statistician in that so much useful and diverse information is contained between the covers of one 290 page book.

REFERENCES

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley, 1958.

Harmon, H. H. *Modern Factor Analysis*. Chicago: University of Chicago Press, 1960.

Rao, C. R. *Linear Statistical Inferences and Its Applications*. New York: John Wiley, 1965.

Thurstone, L. L. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.

George H. Dunteman
Regional Rehabilitation Research Institute
University of Florida

Applications of Electronic Data Processing Methods in Education by John I. Goodlad, John G. Caffrey, John F. O'Toole, Jr., and Louise L. Tyler, assisted by Fred L. Converse and Arthur N. Thayer. A study report supported by the Cooperative Research Program of the United States Office of Education, Department of Health, Education, and Welfare. Project No. F-026, (Los Angeles: Dept. of Education, University of California at Los Angeles, 1965). Pp. 105.

The educator who is interested in the application of electronic data processing (EDP) procedures in an educational setting has to begin his quest for knowledge somewhere. The monograph *Applications of Electronic Data Processing Methods in Education* by Goodlad and his associates represents one such starting point. This monograph is a report of a conference of EDP specialists and an attempt to describe the state of EDP as it existed circa 1964.

The monograph is one of very few publications with which this writer is acquainted that presupposes almost no knowledge whatsoever about EDP and requires no mathematical background in order to follow. As such, it should be welcomed by many educator-laymen (as far as the computer is concerned) whose ability to read technical writing has gone into a positively accelerated decline since their own school days. The monograph begins with a general discussion of education and computer technology. An attempt is made to relate computer technology to the changing American school and to identify possible areas of application of EDP procedures to educational problems.

The second section describes some illustrative applications of EDP and information processing systems in education. The applications range from installations which perform a variety of financial and social bookkeeping operations to experimental projects for computer based instruction and programs for the simulation of an entire school system. Needless to say, the latter applications are still in experimental stages and are not likely to be operational for some time.

The third section of the report presents the results of a conference of EDP specialists and educators with an interest in EDP matters. The conference produced a series of recommendations for the use of EDP in an educational setting and suggested criteria for the funding of projects to explore additional applications of EDP in education. The tone of the recommendations is that our present state of knowledge is quite limited and that a number of studies should be undertaken to explore the ways in which EDP methods can be used in education. The criteria for the funding of projects offer precious little in the way of advice to the grant-seeker or, for that matter, the grant-giver, e.g., "Will the project advance educa-

tional goals by removing a roadblock or otherwise resolving a recognized problem?"

The fourth section sets forth a number of conclusions and recommendations of the authors concerning EDP. The authors seem to have discovered that the computer is a useful tool and could be of significant benefit to education, although they are not quite sure how this is to come about. They are obviously highly impressed with the capabilities of the computer but do realize that technology is no substitute for ideas.

Perhaps the most interesting parts of the monograph are contained in two appendices. One describes the state of EDP in education, while the other presents thumbnail sketches of twenty-seven state, regional, and local EDP installations. The first appendix is a highly readable discussion of the state of the art circa 1964. Some of the material is now out of date, but on the whole, it is the best general statement of how the computer is being used in school settings. Those who have feared that the introduction of the computer into an educational setting heralds a brave new world will find that their fears have no basis in fact—at least so far. Application of EDP methods to date have been limited to essentially accounting and housekeeping functions. In no case, have EDP methods touched the central nervous system of education—the teaching-learning process. This is well documented in the appendix describing reports of operating EDP installations. Summaries of these reports include such basic items as: (1) approximate budgeted expenditures for a fiscal year, (2) hardware used, (3) types of services performed, and (4) size and nature of staff. This reader found the information downright depressing for two reasons. First, EDP methods have been generally used for the most pedestrian tasks of school record keeping. Second, administrative control of EDP installations is, to this reader, too often lodged with business managers and other administrators charged with record keeping functions. It would seem that as long as administrators, whose primary responsibilities are for the keeping of records and financial accounting, are given control of EDP installations, EDP procedures will be primarily used for record keeping and financial accounting. What is distressing is not that EDP methods are being used for such routine administrative tasks, but the fact that currently they are being used for little else.

As can be gathered, this reader was dissatisfied by the small amount of progress that has been made in the use of EDP in educational settings. This is in sharp contrast to the fascinating accounts of EDP applications collected by Borko (1962) and published three years prior to the present monograph. If one were to read the two publications in chronological order, one would get the impression of moving forward by going backward! This is hardly the case.

The above comments are not to be interpreted as being critical of the work of Goodlad and his associates. On the contrary, the authors are to be commended for having exposed such a deplorable situation. It remains, however, that the monograph offers only the barest introduction to the use of EDP methods in education. The knowledgeable person will find little of substance or interest here and will have to look elsewhere for new ideas about EDP applications in education.

Three of the authors of the present monograph (Goodlad, O'Toole, and Tyler) have prepared a book, *Computers and Information Systems in Education,* which is to be published by Harcourt, Brace, and World in September, 1966. Some of the material is drawn from the present monograph.

REFERENCE

Borko, Harold (Editor). *Computer Applications in the Behavioral Sciences.* Englewood Cliffs, New Jersey: Prentice-Hall, 1962.

RICHARD WOLF
University of Southern California

Fortran Programming (II and IV) by L. Dale Harris. Columbus, Ohio: Charles E. Merrill, Inc., 1964. Pp. x + 146.

This is a paperback version of Harris' *Numerical Methods Using Fortran* reviewed previously by this reviewer. The paperback consists of four chapters and the two appendixes from the earlier book. There are some additions in the form of new footnotes and added materials, but usually pages are identical in the two texts.

Thus, the book is designed to be a programming text. The readers who have struggled through the technical manuals and reference manuals of the various computer manufacturers will find Harris a delightful change. The book will be a good teaching document.

Harris, who is an engineer, is not so concerned with storage problems as are many behavioral scientists. For example, no mention is made of chain jobs and temporary tape storage, the use of which the reviewer has found quite helpful. However, the use of COMMON and EQUIVALENCE statements for storage efficiency is adequately explained. The reviewer would prefer to see the storage problem as an added separate chapter. Hopefully, the reviewer's concerns about storage will soon become archaic. But for a while we shall have storage problems; for example, we frequently need ways to squeeze $N \times N$ matrices into storage without needing to reduce the sample size.

Harris' book should become widely accepted as a text for programming classes because of its readability. Competent program-

mers might find the appendixes valuable, but probably will prefer to use the manufacturer's manuals.

W. L. Bashaw
University of Georgia

Numerical Methods Using Fortran by L. Dale Harris. Columbus, Ohio: Charles E. Merrill Books, Inc., 1964. Pp. xii + 244.

Harris attempts "to provide a suitable marriage" of computer programming and numerical methods. The behavioral scientist might be tempted to ignore his work, since it is directed to engineering and the natural sciences. However, the reviewer believes that the computer-oriented behavioral researcher will find the document quite valuable.

The mathematics in the text becomes quite rigorous; however, a non-mathematician who has at least trigonometry and algebra training and who wishes to apply numerical methods without necessarily working out proofs should be able to comprehend Harris easily enough. For those mathematicians who believe that they must prove every theorem that they use, Harris provides sufficient outlines for such proofs.

Most of the book is devoted to the teaching of computer programming in Fortran II and Fortran IV languages. As a programming text, the book is excellent, although for this purpose an instructor might prefer Harris' shorter paperback. (See the review of Harris' *Fortran Programming II and IV* by this reviewer.)

Harris attempts to teach basic programming concepts in his first three chapters plus his fifth chapter. Of special interest is his chapter on iteration (Chapter 3). By treating indexing and DO loops in the context of the numerical method of iteration, Harris teaches the programming techniques and also shows the power of iteration as a numerical tool. He reserves many technical considerations for two large appendixes which should help the student focus on basics.

The remaining chapters deal with numerical analysis techniques. The topics are interpolation (Chapter 4), differentiation and integration (Chapter 6), solution of ordinary differential equations (Chapter 7), finding roots to equations (Chapter 8), and the simultaneous solution of systems of linear equations (Chapter 9). Several solutions are provided for each of these problem areas. Example Fortran programs are presented for each solution.

Perhaps the most enlightening aspect of the text is to show the power of the digital computer to solve complex problems through using fairly naive methods such as trial-and-error. The applied scientist frequently needs to be shown that analytic solutions are not always necessary when numeric solutions can be worked out by a computer quickly, if not always efficiently.

The behavioral scientist will find no discussion of statistical problems and their solutions. However, included techniques have definite applications to the solution of some simulation and Monte Carlo problems, and obvious applications to some of the sticky statistical problems that arise from time-to-time.

In summary, Harris' text will have its largest audience in the natural sciences. The behavioral-scientist-programmer will have no need for much of the book, although the learner or beginner will find it well-written and highly understandable.

W. L. BASHAW
University of Georgia

The Proceedings of the Ninth College and University Machine Records Conference by Don Morrison (General Editor). Monograph 4, Educational Systems Corporation, 1964. (No place of publication is indicated on the Monograph.)

This paper bound volume contains all but three of the papers presented at the annual Machine Records Conference, held at Texas A & M University in 1964. There are a few general papers, beginning with the keynote address by Jack Woolf, dealing with data processing philosophy. Other papers concentrate on more detailed subjects, but again in a generalized sense, such as library circulations, information retrieval, and "Where to—By 1984." The other papers, and by and large the majority, present data processing "experiences" in certain selected and discrete areas, such as data transmission, accounting, budget, accounts receivable, payroll, and class scheduling.

The College and University Machines Record Conference should be praised for covering, quite competently, some aspects of data processing in education; however, the reviewer must take issue with the resultant document. The book is an indication of the present state of the art, and by including certain topics and excluding others is also a statement of philosophy of educational data processors. The areas covered represent the "business" aspects of education. Although there are some comments on instruction, i.e. teaching data processing, or teaching by computer, these topics are not considered in very much detail, nor do they seem to be included as valid areas to be discussed at this conference. It seems strange, and probably remiss, to ignore that part of education exemplified by student learning. There is a paper presented on "Bid-Matching for Panhellenic Associations," but none on the use of data processing for the benefit of instruction. There are a few sentences referring to testing, but no systematic treatment about evaluation procedures for educational product control. There are discussions about audit trails in the business and management areas of educational institu-

tions, but there is a complete lack of discussion of audit trails for student learning. The topics which are covered are well done, but there seem to be surprising omissions.

Several statements were found about which argument can be raised. For example: "teaching by computer . . . can only be used for factual subjects. Interpretations, opinions, and theory cannot be taught by the computer" (p. 43); the presumption on page eight that improvement in instruction results from increasing faculty salaries; the decision that students should have choice of class section (p. 237); and finally, "The advent of serious interest in data processing in public schools is relatively recent" (p. 162). The definition of "recent" seems questionable, since the author of this critique wrote "Automation in Education," published in the *Journal of Machine Accounting* in 1958!

The strong points in the papers include the prevention of duplication of data collections, fewer forms for students to fill out, more rapid collection and distribution of data, and the increased ability to obtain data prior to decision making. The papers fail, however, in fulfilling the hope presented in the opening address, i.e. "Educational objectives, values, and philosophies must take precedence . . ." (p. 2). The papers do not fulfill this obligation.

Richard E. Spencer
University of Illinois

The Prediction of Academic Success by D. E. Lavin. New York: Russell Sage Foundation, 1965. Pp. 182. $4.00.

This theoretical analysis and review of research by a sociology professor at the University of Pennsylvania attempts to pull together and give organizational structure to what is known about the achievement of students at all educational levels. The author begins with a very helpful discussion of the various criteria of academic performance, and goes on to analyze several methodological problems in the prediction of academic achievement such as standardized prediction measures and the interpretation of relationships between predictors and performance. For example, in discussing the problem of over- and underachievement, he points out in a clarifying manner that the intelligence test is not a perfect predictor:

> "It would be more accurate to say that for the prediction of academic performance, ability is but one kind of necessary information. From this point of view, what is left after ability has been used as a predictor is not over- and underachievement, but unexplained variation, much of which may be accounted for by

other predictive factors. In short these terms actually refer to the inaccuracy involved in predicting academic performance from ability measures alone." (p. 25)

After some excellent diagrams and figures illustrating the types of research studies, findings are presented in four broad categories: intelligence and ability factors, personality characteristics, sociological determinants, and socio-psychological factors. On page 107, in summing up cognitive style, the author attempts "an intuitive factor analysis" which comes up with six variables associated with academic performance including social maturity, emotional stability, achievement motivation, cognitive style, achievement via conformance, and achievement via independence. Had the author researched a little more carefully, he would have discovered that the reviewer anticipated him by five years by coming to the same conclusions as a result of an even more extensive analysis in the *Journal of Counseling Psychology* (Summer, 1960). As the author acknowledged, the book is an outcome of a report presented to the Foundation in 1959, and there are few citations of later date. Five years of lag time in a report which attempts a global summary of this area is excessive. Certainly this delay detracts from the book's value.

The book is to be praised in providing some structure and order for the welter of studies in the achievement area. It would have been helpful if the bibliographic section could have been arranged in an order which would make it more possible to look up studies alphabetically. The author is usually very careful about reporting research results accurately. In reporting the Getzels and Jackson study (p. 141), however, he makes the error of stating that teachers preferred the high intelligence to the high creativity students, whereas a more accurate reporting of the findings would have stated that whereas the teachers preferred the more intelligent students over the generality by a significant amount, they did not so prefer the creative ones.

The reviewer agrees with the dust jacket which avers: "As a critical guide to a significant portion of the literature, *The Prediction of Academic Performance* will be a highly important reference work. . . . Admissions personnel, guidance counselors, and school psychologists and administrators will find it useful. . . ." With a little more effort, it could have been much more useful, but it represents a fair, if delayed, start in an area which has cried for straightening out for a long time. What is needed now is for someone else to bring us up to date.

JOHN C. GOWAN
San Fernando Valley State College

Decisions for Tomorrow, Plans of High School Seniors After Graduation by Ralph F. Berdie and Alpert B. Hood. Minneapolis, Minnesota: University of Minnesota Press, 1965. Pp. 185.

Ralph Berdie and Albert Hood surveyed all graduating seniors in Minnesota High Schools in 1961 to identify their post-high school plans and the conditions related to these plans. This study was both a repetition and an extension of the 1950 Minnesota study carried out by Ralph Berdie. The 1961 Minnesota survey and analysis of the data obtained from high school seniors are documented and are described most carefully by the authors.

As a comparative study of 1961 against 1950, a number of questions, as major objectives of the survey, were posed such as: (1) Did the proportions of after-high school pursuits change during the decade? (2) To what extent did proportions of high ability students making these plans change from 1950 to 1961? (3) What are the relationships between ability and these plans? (4) What are the relationships of socioeconomic, cultural, and family conditions to students' post-high school plans, and how have these relationships changed over the eleven years? (5) What are the relationships between personal attitudes and values and the after-graduation plans of high school seniors?

The last question posed marked the most significant change in the 1961 questionnaire from the one used in 1950. Twenty-five items aimed at obtaining information with regard to attitudes and values were added to the questionnaire.

The only finding of any significance for these factors seemed to be that college-bound students in comparison with others expressed more ease in social situations, less difficulty with authority figures, and more favorable relationships with the family. Thus, "attitudes elicited by the questions were directly related to social behavior and social conformity."

The follow-up study one year after graduation showed that the percentage of seniors actually following their plans was higher in 1961 (67 per cent) than in 1950 (64 per cent). The major emphasis of the study was on a particular post-high school goal, attending college. Indeed, 41 per cent of the 44,756 students in the 1961 survey, planned to go to college.

As the study shows, far more high school seniors now than formerly go to college proportionately and in absolute numbers. In fact, the ten year 1950-1960 growth of students beginning college was four times greater than the increase expected from population growth. The authors wisely state, "Meeting the needs of these increased numbers of young people seeking training will be one of the major problems confronting American college educators during the next decade."

The useful generalization about college population growth is firmly buttressed by the research done by the authors, specifically for Minnesota. They summarize their findings in Chapter 2, and in addition they suggest implications not derived from the data analyzed. They found that financial assistance was less of a factor for a college goal in 1961 than in 1950. However, they state the implication that some groups might choose college if financial aid were of substantial amount and if its assurance were known before high school graduation.

There are a total of eighteen chapters plus a foreword and an index. The chapters are:

1. Attitudes, Behaviors, and Plans of High School Seniors.
2. An Overview, Results and Implications.
3. National and Community Changes.
4. Two Surveys Eleven Years Apart.
5. A Follow-up One Year After Graduation.
6. Ability and College Attendance.
7. Effects of School and Community on Decisions For College.
8. Personal Values and Attitudes.
9. Socioeconomic and Ability Variables.
10. Students Bound For College.
11. Students Seeking Jobs.
12. Plans of High-Ability Students.
13. High-Ability Students from Workmen's Homes.
14. Girls Who Planned To Enter Nursing.
15. Girls Who Planned To Attend Business College.
16. Boys Who Planned To Attend Trade School.
17. Boys Planning To Enter Military Service.
18. Profiles: Who Chooses What Plan?

While *Decisions For Tomorrow* is a careful and thoughtful recitation of the data analysis of a comprehensive survey, it should prove useful to a high school counselor when he considers the many factors which affect a student's career plans. The statistical analysis employed in the study will certainly be of help to other educational researchers who, hopefully, will undertake similar studies.

When such studies are undertaken, this reviewer hopes that instruments will be available for greater penetration into the cultural aspects of an adolescent's background since the main items in the 1961 Minnesota study—how many books in the home, what magazine subscriptions—are too limited.

C. Russell de Burlo, Jr.
Tufts University

The Dropout: Causes and Cures by Lucius F. Cervantes. Ann Arbor, Michigan. University of Michigan Press, 1965. Pp. viii + 244. $5.95.

Communicating the findings of a scientific opus to educators, behavioral scientists, and civic minded laymen is a formidable task. Any endeavor to reach such a diverse audience through a single volume is to be lauded. Cervantes undertook such a venture and succeeded in recording and in interpreting the different phenomenal worlds of the high school dropout and of the graduate.

The search for the etiology of early school leaving has swelled in recent years. No matter what the dependent variable has been, most researchers, have found differences between the early school leaver and the graduate. The network of suspected causes includes low intelligence, low reading ability, inadequate personality adjustment, and low socioeconomic status. So many variables have turned up, though, that distinguishing cause and correlate from epiphenomenon has been no simple task. In bold contrast to the hunt and peck research is Cervantes book. The hypotheses guiding the study were derived from theory. The generalized hypothesis that dropouts and graduates are from different family structures was derived from sociological theory of the family. Among the predicted differences were those concerning degree of primary orientation, paternal influence, and involvement with friend families.

To test his hypothesis, Cervantes matched 150 dropouts with 150 graduates according to sex, age, IQ, school attended, and socioeconomic status. The subjects were from New Orleans, Boston, St. Louis, Denver, Omaha, and Los Angeles. Each subject was interviewed for 35 minutes and was, then, asked to answer a questionnaire. In addition subjects from New Orleans and from Boston were given the Thematic Apperception Test (TAT). The youth data were rated by three social scientists. Most tests of the null hypotheses were by the Kolmogorov-Smirnov two-sample one-tail test.

The data gathered indicate that the hypotheses were in the predicted direction. Among the differences observed were those with respect to deferred gratification patterns, youth culture, sexuality, fantasies of aggression, school involvement and membership, facility in communicating, and adequate families. Intensive interviews with a school principal from New Orleans and with a youth counselor from Hollywood, which give spice to the account, are introduced as evidence in further support of the general hypothesis.

This reviewer found Cervantes' book thought-provoking. Despite the highly statistically significant differences between the ratings of the dropouts and the graduates, the conclusions reached by the author are acceptable only as hypotheses. The strength of any set

of conclusions rests upon the representativeness of the samples, upon the adequacy of the procedure, and upon the validity and reliability of the data. The description of the method of selecting the sample is anemic. There is no way to know how ages, IQ's, and other variables had been distributed. IQ's were obtained from school records. Which tests and how recently they were given, vital to careful matching, were not identified. Seventy per cent of both samples were from working class homes. Yet, how socio-economic status of the home was decided was not described. Was Cervantes justified in merging the samples from the six cities? No tests of the homogeneity of the sample were reported. Such tests are crucial. Were the dropout to be a product of a variable correlated with the family organization but not a family variable at all, then the null hypothesis is being rejected when it is true. That intercity differences were observed in family kindred suggests that *city* as a source of variation deserves greater scrutiny. Finally, a substantial proportion of dropouts comes from rural America. These were not considered in the sample. Yet, there may exist differences between urban and rural dropouts.

The dropouts, according to Cervantes, had difficulty in answering the questionnaire. In addition, interviews with dropouts and graduates were under markedly different circumstances. Developing inferences about etiology from *post hoc* interviews is suspect. What criteria the raters used, moreover, is not known. Nor is there any evidence that raters agreed with one another. The author acknowledges the lack of refinement of the nonparametric test he used. Yet his findings read as if family correlated parameters of school holding power were discovered.

The stated focus of the study was "the enveloping social system of the respondent as perceived and internalized by himself." How may the reader be certain that the views expressed represent faithfully the existential past or present? Does he know, moreover, whether the views presented would endure over time? Many dropouts were out of work or in uncomfortable circumstances; many graduates were in favorable circumstances. Perhaps the perceptions of the past were determined by these immediate circumstances and, hence, may only be ephemeral. A more parsimonious statement of the findings would be that the dropouts differ from graduates in *perceptions* of family. This is not the same, however, as saying that the *families* of the two groups differ. When this distinction is made it makes it easier to understand two observations which baffled this reader: (1) parents of dropouts do not value schooling, (2) dropouts ignore or reject parent values. While it may be argued that valuing is not a generalized attribute, worthy of consideration is the hypothesis that a "halo effect" for negativistic reporting operated during the interviews and the data represent these negative

feelings of the dropout. The TAT is known to be sensitive to current environmental presses. The differences in aggression in the TAT stories observed by Cervantes, then, may also be triggered by the frustration or joys of the current circumstances.

Dropouts and graduates, according to the author, represent polar opposites on several dimensions, *e.g.*, troubled-calm; hostile-friendly; pawn-master; sensate-idealistic. This either/or conclusion may result, however, from an artifact of measurement; namely, protocol ratings were in nominal scales. One final point concerns the *ex cathedra* conclusions. Counseling, multi-track curricula, promotion of apprenticeship training, and utilization of resources similar to those under the Economic Opportunities Act were among the suggestions made to promote better use of human resources. While they were not the subjects of the investigation, they nevertheless, offer counsel to the schools, the government, and the community. They deserve investigation in their own right, but not endorsement.

The limitations of the study are cited to offer a context to the reading of the book. The book is recommended to the intended audience although its presumptuous title should be abjured. It will be enjoyed because the dialogue is scholarly without being pedantic and because the data are yeasty without being capricious. The book dispels many popular beliefs concerning the dropout. It showers the reader with new and poignant questions about education and social class. Above all, the book marks an important step in bridling the dropout, at least in research.

Norman M. Chansky
North Carolina State University, Raleigh

Task Experience as a Source of Attitudes by Paul E. Breer and Edwin A. Locke. Homewood, Illinois: The Dorsey Press, 1965. Pp. x + 280. $7.95.

Breer and Locke rigidly delimited their arena, explored it rigorously while continuously refining their design and their tools through replication. Their theory, which they sought to test in a delimited way, is simply that the beliefs and values one has with respect to the world around him have their roots in task experiences.

Though the impact of disparate, delimited, experimental interventions on the value and attitude system of the subjects is highly problematic in this reviewer's opinion, nonetheless one cannot but be impressed with the imaginativeness, thoroughness, forthrightness, and central focus of the seven studies reported here. A few quotes might help to support this impression.

"In two of our previous studies we had put the scales most likely to change at the beginning of the questionnaire (see Chapters 5

and 6). This left open the possibility that the initial scales had evoked some sort of generalized response set. . . . With this possibility in mind, we deliberately put our new neighborhood and family scales first on the assumption they were less likely to change than any of the others in the questionnaire." (p. 218).

"Once the results were in, (changes in subjects' belief in God were in the opposite direction than had been predicted) it was relatively easy to think of reasons why it should have worked this way. In fact, it soon got to the point where we found it hard to explain why we had ever predicted the opposite in the first place." (p. 174).

"There were several considerations involved in planning this final experiment. First of all, we were anxious to do a single study which would incorporate all the major elements in our theoretical scheme. These include behavior, situationally specific orientations, and the two kinds of generalization (lateral and vertical)." (p. 214).

"This was one innovation, i.e., using different tasks for the two different treatment groups. In addition to this change, we decided to use more than one task in each of the two treatments. . . . The reason for the change was simple. We felt intuitively that a task experience encompassing a variety of similar tasks would be more likely to lead to attitude change than an experience based upon a single task, as in our previous experiments. . . ." (p. 140).

In reading this book, the reviewer kept on harking back to an early paper by LaPiere (1934) which indicates (substantiated in subsequent researches) that pencil and paper assessments of attitudes have little or no relevance to behaviors displayed in real life settings. Breer and Locke's basic assumption in the testing of their theory of the relatedness of attitudes to task experience seems to be that the questionnaire responses by subjects regarding their attitudes validly represent the attitudes they evidence in natural settings. This reviewer has grave doubts on this score. He also has grave reservations as to the impact of a delimited, experimental four hour session involving experiences with several laboratory-type task experiences on significant values held by subjects in relation to religion and God, group versus individual effort, and equalitarianism and authoritarianism. The authors, however, present statistical data which indicate such experimental sessions do indeed have an impact on value systems of the subjects and the impact-influence lasts—as measured by a questionnaire—at least more than a week or two.

The theory propounded and to some extent tested in the seven studies (one in a naturalistic and six in a laboratory setting) re-

ported in this book is that an individual's preferences, ideas, and values are derived in large measure from working on tasks and that these cathectic, cognitive, and evaluation orientations are transferred to educational, religious, recreational, and other settings through both lateral and vertical generalization effect. Task is defined (p. 17) as "a piece of work to be accomplished." This definition, however, seems to ignore a number of "tasks" e.g., taking the family on a picnic, using the cards in one's hand to make as many points as possible, hitting a golf ball 200 yards down the fairway, and scaling a rock-face that have other purposes than accomplishing a "piece of work." The bifurcated psyche (pleasure and affection) and socio (task-based) motivation of human actions and interactions hypothesized by Helen Hall Jennings (1949) and Herbert T. Thelen (1960) is ignored or denied by the point of view expressed in Breer and Locke's book.

The authors discourse with skill and cogency on the issue of the types of generality (pp. 256–265) which an adequate theory should demonstrate, the contaminating influence (pp. 262–263) of the demand characteristics of the laboratory situation, and the use of contrast in experimental design (p. 261). However, the generalizability of the Breer and Locke theory regarding how attitudes are developed is called into question for this reader by the authors themselves when they assert (p. 19) that "children take on the beliefs and values of their parents." This apparently is without benefit of task experiences. If it can occur with young persons, why cannot attitude development similarly occur with older ones? The assertion (p. 173) that "The experiment was designed to change subjects' belief in God, and this is precisely what did happen" seems hard to swallow. The independent variables impinging on the subjects to achieve this "change of belief in God" comprised: pre-post questionnaire administrations over four hours of time; the "tasks" of solving jig-saw, miniature tower building, prediction and matrix problems in either an individual or group context; and being paid between $6.00 and $9.00 for the time and effort invested.

The assertion (p. 162) "When effort fails to pay off, the individual will respond by de-emphasizing its importance" seems to ignore the Festinger dissonance hypotheses and research.

Finally (p. 266),

> "To study the relationship between job experience and attitude we first need a measure of each. Developing attitude scales to cover a wide variety of attitude dimension is a time-consuming operation but one that we know can be done."

Have not the vocational interest blank devices and their offshoots made a beginning on this matter long since?

Despite these seeming lapses, the authors retain their forthright and focused stance in pointing out: ". . . it remains questionable to what degree these tasks are representative of the broader population of tasks found in less contrived settings" (p. 260); that, having left the laboratory, they will "be hard put to demonstrate that job experience is responsible for attitude formation" and "that attitude change in the direction predicted by our theory could be explained in terms of socialization" (p. 269). However, they reaffirm their concern in attitude change that is not dependent on verbal communication but that which can be attributed "to the task experience itself." To test and explicate this position, they emphasize that a way has to be found to keep task experience and persuasion separate. Their conclusion is: "There would appear to be no simple solution that is applicable in all situations."

REFERENCES

Jennings, Helen H. Leadership and Sociometric Choice. *Sociometry,* 1949, 10, 32–39.

La Piere, Richard T. Attitudes vs. Actions. *Social Forces,* 1934, 13, 230–237.

Thelen, Herbert A. *Education and the Human Quest.* New York: Harper Bros., 1960.

JOHN WITHALL
The Pennsylvania State University

The Social Science of Organizations: Four Perspectives by Harold J. Leavitt (Editor). Englewood Cliffs, N. J.: Prentice-Hall, 1963. Pp. ix + 182.

This book consists of papers prepared by four of the 24 participants at the two-week Seminar on the Social Science of Organization held at the University of Pittsburgh and supported by the Ford Foundation.

The participants were divided into four smaller groups, each working on a major organizational problem. One member of each group was commissioned to prepare a perspective on the problem which would represent his initial position sharpened and tempered by his two weeks of interaction with persons of other views. Thus, the four perspectives were never intended to be balanced, exhaustive, or reportorial in nature, but they are incisive and competent, and they do reveal the multidisciplinary nature of organizational social science. Of the four authors, George B. Strother is a psychologist; George Strauss, an applied anthropologist; Henry A. Latané, an economist; and David Mechanic, a sociologist.

Strother discusses some of the problems in the development of a social science of organization by beginning with an excellent his-

torical overview. Strother identifies five perspectives or models of organization: ethical models, models which were used by Plato and which have survived Machiavelli to continue to the present; interrelational models, used in the Hawthorne studies of the 1920's and in the more recent laboratory studies of social psychologists with small groups; functional models, which have found sympathy with sociologists, and which define the problem in terms of group purpose and the unit in terms of the aggregate of individuals constituting the organization; structural models, either in the rational form of Weber's ideal bureaucracy or in the classical form of line and staff and departmentation; and mathematical models, associated with the names of Simon, Von Neumann and Morgenstern, and others. Strother then proceeds to a delineation of the logical problems of a social science of organization through emphasizing the specification of unit (individual, group) and of level of abstraction (concept, relation, theory). He concludes by underscoring the current multidisciplinary trend in this area.

Strauss considers certain aspects of power-equalization, the reduction in power differential between supervisor and subordinate. The author criticizes what he refers to as the personality-versus-organization hypothesis, the hypothesis that human motivation in industry is incompatible with the objectives of the organization. According to this hypothesis, once man's physical needs are satisfied, his needs associated with creativity, individual expression, and autonomy come to the fore. In the fulfillment of its routine tasks, the organization can force subversion of these needs, or it can adopt power-equalization techniques with a view toward focussing these needs in the direction of organizational objectives. However, Strauss, noting that the hypothesis has been advanced by academicians, for whom self-actualization needs are strong, doubts the universality of these needs. By questioning the validity of the personality-versus-organization hypothesis, Strauss questions the desirability of power-equalization as a general means for increasing productivity.

Latané considers the rationality model in organizational decision-making. He notes that the model has come under attack on two separate grounds: (1) the model does not deal adequately with real-life limitations on informational and computational ability, and (2) the model assumes maximizing behavior on the part of decision-makers. Latané counters both criticisms by suggesting that rationality models can be built by using generalized payoff matrices stated in terms of some measure of value to be maximized, and that such matrices, normally complex because of the interaction of hierarchical goals and cumulative effects from successive decisions, can be simplified in ways which do not reduce their generality. This may be, in which case the rationality model would be an adequate ana-

lytic model, though this fact is quite unrelated to its adequacy as a simulation model. Latané considers a model of the ideal, not a model of reality, and this distinction does not emerge clearly in his chapter.

Mechanic concentrates on the methodology of organizational studies. In large part, the author organizes methodological problems in terms of the great debate between those two perennial antagonists, the theorist and the empiricist. On the one hand is the theorist with his emphasis on validity, his interest in substantive problems, his penchant for global studies, and his use of variables defined in terms of their practical or theoretical significance; in the other corner is the empiricist, who concentrates on reliability, systematization, restricted studies, and variables defined by the data, perhaps through factor analysis. Regarding research strategy, Mechanic favors a multidisciplinary strategy, which implies different viewpoints (psychological, sociological, economic) of the same organizational problem, rather than an interdisciplinary strategy, which implies a coalescence of ideas and, possibly, an eclecticism which may be too tolerant of weak ideas. Mechanic's chapter is marred only by his use of such terms as "accept the null hypothesis" rather than "fail to reject the null hypothesis," and "negative results" rather than "inconclusive results."

Each of these four perspectives bears the mark of its author's discipline, and their existence within a single cover emphasizes the extent to which the organization has become a center of multidisciplinary study.

EDWARD LEVONIAN
University of California, Los Angeles

The Causes and Cures of Neurosis by H. J. Eysenck and S. Rachman. San Diego, Calif.: Robert R. Knapp, 1965. Pp. xii + 318. $7.95.

As stated in the introduction, "This book is an introduction to post-Freudian methods of diagnosing and treating neurotics." If this sentence is taken as a statement of purpose, the contents of the book broadly overshoot the authors' goals, and, in other ways, fall far short of surveying the myriad of theories and treatments which have been applied to human beings debilitated by emotional and behavioral disorders in the post-Freudian era.

The subtitle, "An introduction to modern behaviour therapy based on learning theory and the principles of conditioning," gives a clearer statement of purpose and content than either the title or introductory statement. As such, this book joins Eysenck's two anthologies and the recent books by Wolpe, Bachrach, Franks, Wolpe/Salter/Reyna, Ullmann/Krasner, Krasner/Ullmann, and

Bandura/Walters in expressing not only dissatisfaction with the scientific status and effectiveness of traditional "depth" theories and therapies, but a more promising alternative—the principles and techniques of learning.

To the methods of treatment derived from the principles of learning, Eysenck has previously given the generic label, "behaviour therapy," to distinguish them from "psycho-therapy," and especially, psychoanalysis. This is perhaps an unfortunate term, because when accompanied by such statements as, "Get rid of the symptom (skeletal and autonomic) and you have eliminated the neurosis," the term itself tends to suggest that the "behavior therapist" is interested only in minor, specific, observable motoric behaviors of a superficial sort; that the client may be dehumanized, mechanized, or otherwise, and; that the application of these methods still operates within a disease-analogy model of symptom-with-underlying-unconscious-cause. For the reader who is not troubled with such implications, or who overcomes his emotional reactions for sufficient time to read the entire text, these inferences will be largely dispelled, *e.g.* "behaviour" encompasses not only motoric behavior, but also autonomic or emotional behavior, and cognitive or ideational behavior; such distinctly human problems as social anxiety, obsessive thoughts, and hysteria are functionally analyzed and treated; term such as "cure" and "symptom" are primarily used for alliterative convenience, and certainly do *not* allude to an acceptance of a "disease model," even by analogy. While an extended discussion directed at the above questions would have enhanced the comprehension of this book, an excellent discussion does appear in Ullmann and Krasner's *Case Studies in Behavior Modification*, which should be considered as a companion volume by prospective users.

In undertaking an "introductory" text in this area, these prolific and provocative British authors have managed to squeeze massive amounts of data, theory, ideas, and an extensive bibliography to 1964, into the 318 pages of text. Although the theoretical chapters are admittedly biased, they provide the most succinct summary currently available of the research and theory stemming from Eysenck and his followers. These summary chapters provide a basic reference for relating diagnosis and treatment of various disorders included later in the text; but since they also go beyond the scope of the authors' stated purpose, they may perhaps contribute to some difficulty with general organization.

The book opens, quite naturally, with a discussion of the problems of classification and the nature of neurosis, and closes with two chapters on the results of "behaviour therapy,"—refreshingly, with data. In between are included chapters on factor analytic studies of personality, drugs, and heredity. These are followed by

the real "meat" of the book, which consists of a review and commentary on the various methods and techniques of "behaviour therapy" and the wide range of problems to which they have been applied. The difficulty in organization apparent in these chapters seems to stem from the inadequacy of current nomenclature in handling the variety of problems which human beings can find for themselves—and the lack of anything better to replace the current system than a broad two-factor dimensional classification. Thus, we find several chapters which are problem centered, organized around more-or-less traditional diagnostic categories, such as "Anxiety States" and "Hysterical Disorders," with the variety of treatment methods which have been applied to them reported within each chapter. Then, a chapter devoted solely to "Avoidance Conditioning and Aversion Treatment," with the variety of disorders so treated reported within the chapter. This is followed by a chapter on "behaviour therapy" with psychotics, and then three chapters on "Children's Disorders" organized still differently. The difficulty in organization, however, should not overshadow the excellent content included within the chapters on treatment. They do provide a most compact overview of the published literature on the range of problems and types of treatment programs which have been attempted.

The theory and assessment chapters, on the other hand, will insure that Eysenck maintains his preferred role as a controversial author. In the first table, for example, the list of ten "most important differences between psychotherapy and behaviour therapy" is likely to raise the hackles of most "behaviour therapists" as well as a few "psychotherapists." In addition, the chapters on "Dimensions of Personality" and "The Biological Basis of Personality" present data from factor analytic and twin studies to support Eysenck's theories on extraversion-neuroticism-psychoticism in such a way as to impress the naive reader that there is no question but that these are "truth." Perhaps the authors' most bothersome tendency in the theoretical chapters is to make some rather broad leaps between data and conclusions by means of semantic equalities, *e.g.*, Pavlovian "excitation," attributed to introverts, is equated with cortical "activation" or physiological "arousal." Data on the relationship between "drive" and performance are then taken as support for introversion-extroversion hypotheses.

In short, this book provides, in a single cover, a concise summarization of the research and theories of Eysenck and his followers across a broad range of areas—not merely "neurosis." Work on behavior disorders to 1964, from a learning approach, is comprehensively described and supported with available data relating both to outcome and theory. The chapters on children's disorders, especially, appear to be the most comprehensive to date, and the

extensive bibliography should aid interested readers in selecting original source materials. Since the book is written as an introductory text, it should not be considered either as a manual on technique or as a book on methodology. Where previous training is sufficient to allow critical examination, it would seem suited to give a broad overview for introductory courses at the graduate or postgraduate level. The optimum use of this book would call for supplement by additional texts emphasizing methodology, and a more thorough examination of the theoretical assumptions of the "causes" of behavioral disorders within a social framework.

GORDON L. PAUL
University of Illinois, Urbana

Individual Differences in the Classroom by R. Murray Thomas and Shirley M. Thomas. New York: David McKay Company, Inc., 1965. Pp. v + 567.

The complexities of the human organism are reflected in individual differences which are staggering to the imagination. Basic to effective classroom instruction is the teacher's ability to cope with these differences. In addressing themselves to this gigantic task the authors state: "The most persistent, discouraging problems faced daily by teachers are usually caused by the fact that pupils differ in so many ways. . . . there seems to be no single modern textbook designed to aid [teachers] adequately with their problems in this realm. The purpose . . . [of this book] is to help fill this gap" (p. 4).

In limiting the scope of their book, the authors elect to focus on (1) intellectual differences, (2) differences in specialized abilities, like those affecting success in music, art, and motor activities, and (3) psychophysical differences, such as variations in hearing, seeing, and speaking. They specifically exclude group differences such as race, religion, and social class.

In developing their major thesis, the writers have attempted to answer several questions that appear vital to classroom instructors.

1. What is the nature or definition of this facet of the child's or youth's life?

2. What range of differences will teachers be likely to find in a typical classroom?

3. How can educators recognize or measure the differences?

4. What kind of classroom organization and, in some instances, school-wide organization can best care for these differences?

5. What specific methods and materials should a teacher use to help students who have these differences learn at their optimum level?

The book is divided into four major parts: The Problem of Differences; Intellectual Differences; Artistic and Motor Differences;

Psychophysical Differences. Part I represents an attempt to build the case for the remainder of the book. Indicative of the lively writing style employed is a statement from Chapter 1: "Children are not created equal, nor do they become more alike as they grow older. Rather, by the time they enter school the inequalities among them—intellectually, physically, and in social behavior—have increased manyfold. As they move upward through the grades, the differences increase even further" (p. 3). Chapter 1, then, casts some light on the extent and importance of differences, whereas Chapter 2 is aimed at enabling teachers to ". . . examine their beliefs about individual differences so that they might form their values into a conscious, consistent system" (p. 34). A further stated purpose is to "introduce a philosophic orientation toward individual differences on which subsequent chapters are based."

Part II deals with various techniques of appraising individual differences, their range, and various administrative and methodological problems encountered. Considered in Part III are those aptitudes which seem to have little, if any, relation to those required in traditional academic pursuits. Included are differences in the graphic arts, those related to music, and those involving other types of "great and fine muscle coordination." Finally, in Part IV, emphasis is placed on psychophysical deviations. Included are vision, hearing, and speech disorders. Also treated are crippling conditions, such as cerebral palsy, epilepsy, heart conditions, disfigurements, malnutrition, and diabetes.

A noteworthy feature of this book is the authors' superb ability to come to grips with highly controversial issues. Instead of avoiding or glossing over such issues, they develop arguments for and then arguments against the issues and finally indicate the preferred courses of action from the standpoint of the classroom teacher. This reviewer was impressed with their ability to make all points of view seem appropriate under given conditions. This agility is demonstrated in their discussion of the current controversy with respect to the single or unitary-factor theory of intelligence, as opposed to the group or multiple-factor theory. "Then what is intelligence really, a single capacity or several clusters of abilities? The truth of the matter, from the standpoint of classroom applications, seems to be a combination of these two theories" (p. 58). Again, in Chapter IV, when the subject of homogeneous and ability grouping is treated, the writers avoid antagonizing groups by giving conditions under which ability grouping works best.

This writer is hard pressed to find major weaknesses in this invaluable contribution to an understanding of the myriad of differences among growing, developing children. Obviously the writers were unable to treat, in depth, many of the problems. It seems that they have succeeded in pulling together in one volume many aspects

of the problem which have been widely scattered throughout the professional literature. They have focused attention upon many areas which tend to be neglected in other sources. This especially applies to so-called psychophysical deviations.

In a book of this kind the writer must choose those problems which seem most important. It is almost impossible to deal with *all* aspects of such a diverse area. This reader did note what he considers to be two rather serious omissions. First, the problem of creativity was totally ignored. It would seem that the writers could have at least cleared up many misconceptions relative to the relationship between academic aptitude and creativity.

Another serious omission is the problem associated with being left-handed. Since almost 10 per cent of our youth must "limp along as best they can in a right-handed world," this reviewer is perplexed as to the reasons for this omission. It should be noted that this problem receives relatively little attention elsewhere.

Another shortcoming of the book seems to be associated with the authors' stated purpose of assisting the teacher in formulating his values into a "conscious, consistent system." While it follows that serious consideration of the problems to which this book is addressed should contribute to such a lofty purpose, this reviewer feels that the task has not been completed. Perhaps a culminating chapter, designed to help the reader make the desired connections, would have been most useful. It has long been known that the transfer value of learning is directly related to the extent to which the instructor teaches for this transfer. Accordingly, it seems incumbent upon the writers to lead the reader in the desired direction.

It should be emphasized that the shortcomings are few and that they apply almost exclusively to areas of omission. The authors have struck at some of the most basic practical problems of teachers. Indeed if this book were read carefully by most school practitioners its impact would be marked. Many professional educators should find the book useful in a variety of teacher-preparation experiences.

KENNETH H. HOOVER
Arizona State University

Identification in Child Rearing by Robert R. Sears, Lucy Rau, and Richard Alpert. Stanford, California: Stanford University Press, 1965.

The focus of the study reported in this book is the interrelationship of child-rearing practices and the behavior of four-year-old children. The major dynamic is the process of identification. Identification is presented first in the classic Freudian fashion, beginning with anaclitic identification and subsequently developing into

defensive identification. The entire process of identification in the child was theorized to be a function of the child-rearing practices used by parents.

An early problem faced by the researchers emerged from the differences between a descriptive theory of personality and a behavioral theory which had predictive potential. It was necessary to translate the Freudian concepts of anaclitic identification and defensive identification from the traditional, descriptive psychoanalytic theory into a dynamic behavioral theory which provided the potential for generating testable hypotheses. The data to test the hypotheses were collected through behavioral observations: through questionnaires to teachers and parents; and through interviews with children, with teachers, and with parents. The analysis of the data was confined primarily to tests of differences between means either through a critical ratio or t and to the computation of correlations among the scales.

The study is in the traditions of the careful field studies of child development in which several of the authors have previously participated. The interpretation is based less upon a carefully delimited mathematical experimental design than upon the application of some statistical models to test certain ideas about field activities in the child-rearing interactions. As a result, in some cases, the interpretation of significance was concerned more with whether any support for the idea could be detected than in terms of a rigorous theory of the reliability of statistics. The authors, at times, are quite willing to conclude that the differences observed in the data favored their analyses, even though the statistics were somewhat short of commonly accepted standards of statistical significance.

In the development of the theory, the authors posited that dependency would be an important factor in the motivational system of the child and from it would emerge the functions of identification. The data did not wholly support their contention. In examining the antecedents of the relationship between child-rearing practices and the development of identification, the authors reached the conclusion that initially, children of both sexes adopt feminine-maternal patterns of behavior. Boys, subsequently, at some point in the first three or four years, begin to alter their behavior towards the traditional masculine role. The activities of the father seemed to be insignificant in the development of the generalized male sex role of the son. Little attention is given to the commonly observed attachment of daughters toward fathers at about the same age. Contrary to traditional theoretical assumptions, if the daughters also evidence a shift toward the masculine role at this time, their task in establishing their final generalized female sex role would be the more difficult. The boys would pass through a two-stage process by identifying first with the maternal role and subsequently shifting

to the masculine or father role. In contrast, the girls, also beginning with primary identification with the maternal role, subsequently would shift toward identification with the male or father role, only to have to shift their identification back to a feminine orientation. Thus, for girls a three-stage process is presented in contrast with the two-stage process of identification for boys.

In general, the notion of anaclitic identification which develops very early in a child's life, received substantiation. In contrast, however, the subsequent defensive identification received little, if any, support. The notion of a generalized trend of defensiveness was not supported.

The book is an outstanding example of a carefully designed and executed field study. It is based upon a series of carefully developed theoretical postulates. Observation and measurement procedures were devised specifically for the problem at hand. The processes of data collection relied heavily upon framed interviews and well-trained observers. The resulting data were probably far more specific and useful to the problem than they would have been had more standardized or traditional measures been used.

The resulting mass of data was almost overwhelming. The last third of the book presents the data-gathering instruments and discusses some of the problems met in collecting and processing the data. A total of 566 variables was coded: of these 246 were parental, and 310 concerned with the child. Some of them proved unacceptable or unusable. The final report concerns itself with 373 variables: 184 parental, 189 child. For most of the variables, the total N used in the correlations was 40. In some cases, because of incomplete data, the N was somewhat smaller.

It is difficult to avoid speculating on whether or not a more carefully developed research design and the application of more exotic statistical technique could have added significantly to the findings in the book. However, at the same time, the converse speculation cannot be avoided. Perhaps the results are more significant because they were gathered under actual field conditions than if the study had been conducted in a more controlled, laboratory setting. To put it in another way, a more coherent design for the generation of data might have detracted as much from as it could have added to a study. Even though a sophisticated design would have been more in keeping with the currently accepted standards for educational research, the ultimate contribution might have been considerably smaller.

The study is an excellent example of an attack upon a tremendously complex problem. The complexities of the reality were not permitted to reduce the care given to the theoretical basis or to the scientific analysis in spite of the fantastic number of variables with which the researchers were forced to deal. They were able to carry

their work through to completion. The study suggests that many of the phenomena which are endemic to the developmental process can be effectively studied on a short term basis. It also demonstrates that a field study need not be conducted at the expense of a theoretical formulation, but can indeed profit from a careful theoretical basis.

GERALD T. KOWITZ
University of Houston

Developmental Patterns in Reactions to Frustration by Udai Norain Pareck. Asia Publishing House (distributed by Taplinger Publishing Co.), 1965. Pp. ix + 182, $5.50.

While the title may lead the unsuspecting reader to assume that he is about to enlarge his understanding of some aspect of child development specific to frustration, he may find himself engrossed by the details of a project to translate an American text for use with children in India. Dr. Pareck, in a doctoral dissertation at the University of Delhi, has developed a children's form of the Picture-Frustration (P-F) Study for use in India and this book is a summary of his efforts in validating this procedure. The problems of altering the P-F illustrations to "Indianize" the situations, the difficulties of translating into Hindi American terms which have no Hindi counterpart, and the tryout procedures are described in detail. While this is not the first effort at importing a text from one culture to another, it is of historical interest because the author does not feel the need to write in the cold, detached manner that characterizes many technical writers.

Whether there is more here than anecdotal material about a text being transformed for use from one culture to another is difficult to say. Saul Rosengweig, the creator of the original P-F Study, tells us in a foreword to the book that the sample of 1,000 children is larger than his original standardization. It is difficult to determine whether the sample is, as the author claims, representative of children in India from ages 4 to 13, rather than representative of children *in school* in Delhi. In view of the extent of illiteracy in India, the standardization group from nine schools may be atypical of Indian children in general. This suspicion of atypicality is increased when it is considered that Hindi, the language of the new version, is spoken by slightly less than half the Indian population (*World Almanac*, 1966, p. 609).

Dr. Pareck makes cautious claims for the validity of his version. Correlations of test scores with behavioral ratings by teachers were nonsignificant and differences between delinquents and the sample were nonsignificant on six of seven scoring categories. The author was more successful in demonstrating shifts in P-F scores as a re-

action to induced frustration and in a "regional analysis" of homogeneity. Comparisons of scores from Japanese and American manuals of the P-F with the author's sample is presented as a study of cultural differences. No data relating to matching of samples on such categories as sex, educational achievement, intelligence level, or socioeconomic level are presented.

Although the comments have tended to be critical of Pareck's efforts, there is much of interest to test constructionists in this book. Not only is the description of the translation process of interest, but his reviews of the research on frustration and its measurement (Chapter 2) and of research with the P-F (Chapter 3) will also prove useful. The overwhelming majority of his references are to American sources and this is, in the reviewer's opinion, the real failure on the part of the author. Here was an opportunity to justify printing and distributing the book in the United States in the first place: a review of what Indian psychologists are doing in personality assessment research.

PHILIP HIMELSTEIN
Texas Western College of the University of Texas

Substrata-factor Reorganization Accompanying Development in Speed and Power of Reading at the Elementary School Level by Harry Singer. Riverside, California: University of California, Riverside, 1965. (Multilithed) Pp. i + 276.

This impressive monograph, a final report of research supported by the U. S. Office of Education over a number of years may be classed as a descriptive, developmental study. This cross-sectional study of pupils' reading behavior in grades three through six is based on substrata-factor theory. The theory attempts to explain the mental structure and dynamics involved in reading ability. The findings from this study, derived by sophisticated statistical techniques, appear to make significant contributions of interest to those concerned with the psychology of reading and with educational and psychological testing.

Design

The research design consisted of (a) the selection of a battery of 28 reliable and valid tests for prediction of speed and power of reading, (b) administration of this battery to a total of 927 pupils in six elementary schools, and (c) statistical tests of the hypotheses by means of substrata analysis, significance of difference between means of known-groups, and principal components factor analysis. Some of the language perception tests that Singer created especially for the study are novel, particularly the test on Matching

Word Sounds, a silent reading test of sounding out whole words. His interesting Syllabication test is designed to measure incidentally application of general principles. The substrata analysis employed, an extension of the Wherry-Doolittle Multiple Test Selection Technique, is a useful tool for investigating and discovering statistically hierarchial structure or psychological organization.

Findings

The major developmental hypothesis of this study was confirmed: A sequential development of a hierarchial organization of substrata-factors did accompany improvement in speed and power of reading. A reorganization of substrata-factors tends to occur in which first level substrata-factors at lower grade levels become subsumed within substrata-factors at higher grade levels. Certain substrata-factors, however, (a) Visual Verbal Meaning for Speed and Power of Reading and (b) Speed of Word Perception for Speed of Reading, are *first* level factors *throughout* the intermediate grades (p. 195). The general premise that speed and power of reading are separate but interrelated components appeared to be tenable at least as early as the third-grade level.

The minor hypothesis concerning a transition or shift in auditory-visual dominance as the average pupil moved through developmental stages in reading was also confirmed. Throughout the intermediate grades, visual processes were dominant for Speed of Reading. But both visual and auditory processes continued to operate in Power of Reading. There was, however, some evidence that auditory abilities and processes may have shifted from more peripheral to more central determinants of Speed and Power of Reading. Concomitant with development of general reading ability, auditory and visual subsystems became more closely integrated and functioned as a more centrally mediated communication system. This integrated subsystem appeared to have an interfacilitory effect upon improvement in speed and power of reading (p. 197). It was inferred that the auding domain entered into speed and power of reading as a more centrally-integrated subsystem as individuals, in general, improved in speed and power of reading (p. 194).

Significant differences between means of known-groups such as boys and girls and fastest, slowest, most powerful and least powerful readers were found. It is likely, however, that all the known-groups at grade levels three through six mobilize at least minimum amounts of word meaning and word perception factors in their general working systems. In grades three through six, within-grade level superiority in Power of Reading was associated with quantitatively more rapid development in almost all substrata selected variables. Within-group superiority among the most powerful readers at the sixth grade lay primarily in the word meaning domain

(Visual Verbal Meaning and Meaning of Affixes) (p. 192). During grades three through six, the slowest and least powerful readers were approximately three grades behind the development of the fastest and most powerful readers. For the benefit of these "deprived" children, Singer developed the concept of the usefulness of substrata diagnosis and evaluation from the statistically determined model to investigate the working system of these children for attaining speed and power, and to plot parsimoniously remedial and developmental action. The teacher could plot and compare graphs for the individual child with psychographs (derived from mean standard scores) for his referent group at each grade level.

Perhaps the most significant finding for those interested in educational and psychological testing is that some variables in reading appear to undergo systematic organizational changes as individuals improve in speed and power of reading. The variables could have, therefore, quantitatively and/or probably qualitatively different validities.

The term, "substrata validity," coined by Singer, has been implicit in substrata analysis, but this monograph makes the neologism explicit. Essentially, substrata validity implies that the validity of a variable is a function of its context within an empirical and theoretical framework. Within this framework, a variable can have a low validity relationship with a criterion several levels removed, but can have a high predictive validity for a subcriterion nearer in the framework. Most importantly, whatever its quantitative relationship, its validity can also be judged on a qualitative criterion of being necessary in the mental organization for reading. To illustrate, visual or auditory acuity shows little correlation with reading comprehension. But, either one is important for discrimination and perception which, in turn are foundations for abstractions and generalizations within a particular modality, reading or auding.

Factor analytic findings were consistent with an observation concerning a typical shift in instructional emphasis toward word meaning and mediational processing systems as the average pupil improves in speed and power of reading. As the grade level rises, word meaning variables do increase their loadings on Factor I, Visual Verbal Comprehension which finally emerges at the sixth-grade level as the combined Audio-Visual Verbal Comprehension factor. An implication for instruction and testing, drawn from systematic differences that were found between visual and auditory word recognition subsystems in relationship to Speed and Power of Reading was the following: objectives in grades three to six might be differentiated with emphasis upon visual word recognition for speed of reading and emphasis upon auditory and analytical subsystems for power of reading (p. 195).

The implications of this study for the development of speed and power in reading in the classroom are the following: (a) In general the working systems uncovered at the successive grade levels provide a basis for determining important subskills and for deciding what relative degree of emphasis should be placed upon them. (b) The developmental models of first level substrata variables give curriculum and test makers an overview of the scope and sequence of factors which allow for individual differences in Speed and Power of Reading, grades three to six. (c) Because there are some factors common to the developmental working systems for Speed and Power of Reading, instructional emphasis on common factors would contribute to the development of both working systems. Subsystems needed for shifts from speed to power might be developed, perhaps by alternating instruction from accuracy to speed. In planning a reading program it appears that other factors —for example, attitudinal, motivational—should be included, since not all the variance was accounted for at any grade level.

In the reviewer's opinion, this research represents a scientific approach to reading instruction that will take the discipline even further away from the "chest thumping, eyelid lifting" mode of reading diagnosis of the past. With the discovery of ways of measuring additional variables that enter into reading and with the accumulation of studies of the type reviewed here, the discipline will move even farther in the direction of furnishing a scientific foundation for diagnostic testing and reading.

SARA W. LUNDSTEEN
University of California, Santa Barbara

Clinical Aspects of Remedial Reading by Clifford J. Kolson and George Kaluger. Springfield, Illinois: Charles C. Thomas, Publisher, 1963. Pp. 146. $5.75.

This very short book purports to be a "complete" program for setting up procedures and practices to help children overcome reading problems. Of its twelve brief chapters that deal with diagnosis of reading difficulties, their treatment, and the reading clinic, the three on diagnosis will be of greatest concern for the reader of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT.

Here, the authors divide reading disabilities into two categories: (1) the *primary* or congenital defects the symptoms of which they say are word blindness, disorientation of letters in spelling and digits in calculation, inability to recognize letter configurations, and right-left disorientation, and (2) the *secondary* or acquired defects the symptoms of which are slowness in synaptic transmission (the passing of nervous impulses from one neuron to another),

visual and hearing defects, poor motor coordination, endocrine imbalance, low intelligence, poor home environment, poor teaching, and emotional problems. None of these so-called primary or secondary defects is spoken of in general terms, however. Rather, a group of specific symptoms are given and these are then contended to be part of a generalized defect. How the several items in the secondary category are "acquired" is not explained.

The approach to diagnosis given in the book is exemplified by the citation of authors whose works are considered a "must": Delacato, Fernald, Herman (sic, see Hermann), Smith and Carrigan, Gillingham; and the centers for clinical internship deemed exceptional: Parker School in Chicago and the Orton Reading Center at Winston-Salem, North Carolina.

The diagnostic procedures suggested follow this pattern. A child with reading disability is first given a two-part screening test. Here he (1) answers nineteen questions such as "Examiner touches subject's left hand. Which hand is this?" "Shows picture of left eye. Which eye is this?" and (2) writes dictation at an uncomfortable speed. If he "fails" these tests (norms for passing not given) he is said to have a "primary" disability, and then is given a mental test, a photographic test of his eye fixations, a silhouette completion test, visual and hearing tests, a figure drawing test (the Bender Gestalt is not used), the Kephart scale (walking a board, ocular pursuit, etc.), the Mills Learning Test, and diagnostic reading tests (a test of speech defects is notably missing).

If the subject "passes" the above screening tests he is considered to have a secondary disability and is given only a mental test and diagnostic reading tests since "It is better, therefore, to dispense with an etiological diagnosis and proceed to a therapeutic diagnosis . . ." because "In the diagnosis of most secondary reading disability cases it is impossible to determine the initiating cause or causes." This is stated in spite of the number of readily ascertainable aspects given as secondary causes.

It can be seen that this book breaks sharply with those descriptions of diagnosis usually found. Especially is this true with the emphasis given to neurological as versus psychological causes of reading disability. In the light of the existing research evidence one must view the premises for diagnosis given here as extremely tentative, since relatively few reports have come out which relate reading problems to neurological defects. The evidence so far is that brain damage is seldom the cause of reading disability. As Arthur Gates says, the concept of word blindness "can rarely, if ever, be applied to reading disability." Likewise, little relationship has been shown between lateral dominance and reading disability. Correlations between reading ability scores and visual tests also tend to be negligible. The relationship between hearing disability and reading, too,

depends on the emphasis given to oral instruction. None of this evidence is examined in the book.

Proof that these and many other research findings are wrong should have been an impressive part of the theory of diagnosis given in this book. Unfortunately, such an emphasis was not the case. Instead, Kolson and Kaluger chose to quote only the relatively few studies that give them some support, and to repress or ignore the weight of evidence against their diagnostic scheme. While this book errs basically in its omissions of evidence, it still is somewhat useful as a challenge to the many more popular theories and practices in the diagnosis of reading difficulties. There is little doubt that such disputations are necessary for progress in this area.

PATRICK GROFF
San Diego State College

ERRATUM

The article "Kuder-Richardson Reliabilities of Classroom Tests" by Edward E. Cureton which appeared in the Spring of 1966 issue of this journal contained an error. σ_t^2 on page 14, five lines above the last should read σ_t.

Dr. Cureton is indebted to Dr. Richard E. Spencer of the University of Illinois for pointing out this error.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 815–824.

ρ_m AS AN "ERROR-FREE" INDEX OF RATER AGREEMENT

JAMES C. NAYLOR AND E. ALLEN SCHENCK
The Ohio State University

THERE is a growing interest in the use of the multiple regression model as an analytical device for describing the strategies of individuals in multiple cue choice situations. For example, the studies by Smedslund (1955), Uhl (1963), Peterson, Hammond and Summers (1965), Summers (1962), and Schenck and Naylor (1965) are all instances of the application of regression analysis to the evaluation of performance strategies in probability learning tasks. Similarly, Naylor and Wherry (1965), Wherry and Naylor (1966) and Madden (1963) have used the multiple regression model to capture the strategies or policies of judges in attitude research.

There is an important distinction, of course, between these two kinds of research investigations. The major distinguishing feature lies in the presence or absence of a criterion of "correct" behavior. In studies of decision-making, concept formation, and/or probability learning, the response of the subject to each stimulus can be evaluated in terms of its correctness, whereas in the case of attitude studies there is no such animal as the "right" response.

In several recent methodological articles (Hursch, Hammond, and Hursch, 1964; Hammond, Hursch and Todd, 1964; Tucker, 1964; and Naylor and Schenck, 1965), a number of performance indices based upon the regression model have been developed for examining the behavior of individuals in those situations where a criterion of correct behavior is present. However, as yet the applicability of these indices to choice behavior situations of a non-criterion nature has not been forthcoming. Our purpose is to relate the model and its performance indices developed in the criterion case to the non-criterion situation.

Model

Consider the typical attitude assessment or rating situation where two judges are asked to evaluate the same set of n stimuli on some defined criterion dimension and where each stimulus has k dimensions which are potentially relevant to the judgment required of the raters. Such a situation is schematized in Figure 1, where

X_i = potentially relevant stimulus dimensions or cue values ($i = 1, 2, \cdots, k$).
Y_A = response of judge A.
Y_B = response of judge B.
ρ_{Ai} = correlations over a population of stimuli between stimulus cue i and the response of judge A.
ρ_{Bi} = correlations over a population of stimuli between stimulus cue i and the response of judge B.
ρ_{ij} = correlations over a population of stimuli between all pairs of k stimulus cues.
β_{Ai} = standard score regression weights for predicting Y_A from X_i values.
β_{Bi} = standard score regression weights for predicting Y_B from X_i values.
Y'_A = predicted response for judge A based upon a least-squares regression equation relating his responses to the stimulus cue values.
Y'_B = predicted response for judge B based upon a least-squares regression equation relating his responses to the stimulus cue values.

Assume $Y_A, Y_B, X_1, \ldots, X_i, \ldots, X_k$, to be random variables with means of zero and variances of one, that is,

$$E(Y_A) = E(Y_B) = E(X_i) = 0$$

and

$$E(Y_A^2) = E(Y_B^2) = E(X_i^2) = 1.$$

Further, define Y'_A and Y'_B to be least-squares estimates of Y_A and Y_B such that

$$Y_A' = \Sigma\beta_{Ai}X_i \quad \text{and} \quad Y_B' = \Sigma\beta_{Bi}X_1. \text{ [1]}$$

[1] On all occasions, summations are over the k cue values.

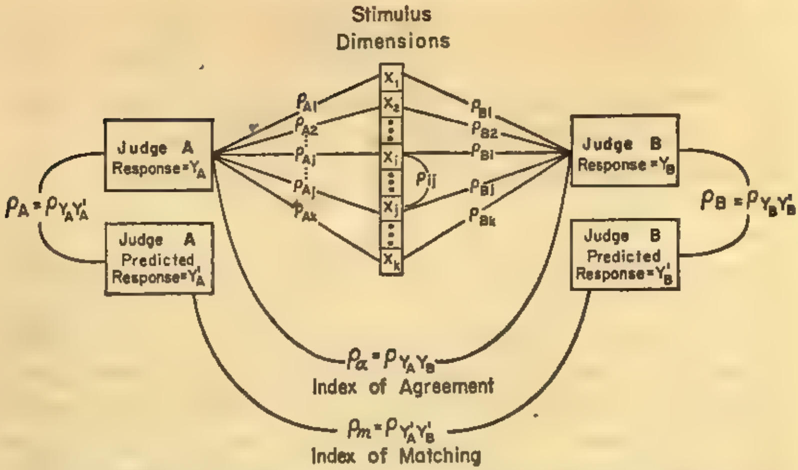


Figure 1. Diagram of lens model showing the relationships among the stimulus cue dimensions and the judges' responses.

Then

$$Y_A = Y_A' + Z_A \quad \text{and} \quad Y_B = Y_B' + Z_B$$

or

$$Y_A = \Sigma\beta_{Ai}X_i + Z_A \quad \text{and} \quad Y_B = \Sigma\beta_{Bi}X_i + Z_B,$$

where

$$E(Z_A) = E(Z_B) = 0 = E(Z_AX_i) = E(Z_BX_i)$$

and

$$\sigma_{Y_A}{}^2 = [\sigma_{Y_A'}{}^2 + \sigma_{Z_A}{}^2] = 1 = [\sigma_{Y_B'}{}^2 + \sigma_{Z_B}{}^2] = \sigma_{Y_B}{}^2.$$

One can then define the correlations in Figure 1 as

$\rho_{ij} = E(X_iX_j) =$ correlation between stimulus cues i and j.

$\rho_{Ai} = E(Y_AX_i) =$ validity of cue i in determining judge A's responses.

$\rho_{Bi} = E(Y_BX_i) =$ validity of cue i in determining judge B's responses.

$$\rho_A = \frac{E(Y_AY_A')}{\sigma_{Y_A'}} = \sqrt{E(Y_AY_A')} = \sqrt{\Sigma\beta_{Ai}\rho_{Ai}}$$

= multiple correlation of judge A, i.e., the degree to which one can predict his responses using all k cues.

$$\rho_B = \frac{E({}_A{}_B Y_B')}{\sigma_{Y_B'}} = \sqrt{E(Y_B Y_B')} = \sqrt{\Sigma \beta_{Bi} \rho_{Bi}}$$

= multiple correlation of judge B, i.e., the degree to which one can predict his responses using all k cues.

The last two coefficients have the well-known property that

$$\rho_A = \sigma_{Y_A'} \quad \text{and} \quad \rho_B = \sigma_{Y_B'}.$$

Agreement between Judges

Let us call the relationship between the responses of two judges (A and B) "response agreement" and define it in the same manner as "achievement" has been defined where the model was used to analyze performance in the presence of a correct response (e.g., see Hursch, Hammond and Hursch, 1964; Naylor and Schenck, 1965); thus

$$\rho_a = E(Y_A Y_B) = \text{response agreement between judges.} \tag{1}$$

The correlation between residual scores is then defined as

$$C = \frac{E(Z_A Z_B)}{\sigma_{Z_A} \sigma_{Z_B}}. \tag{2}$$

Equation (1) is the normal manner in which agreement between judges is defined, i.e., the correlation between their sets of responses. However, we are usually not only interested in the agreement between what the judges actually do, but we are also interested in the agreement of their basic judgmental policies. These are not quite the same thing. Previously we have defined the policy of a judge in terms of "what that judge does" (Naylor and Wherry, 1964). This definition is no longer convenient—rather let us now define the policy of a judge as his least-squares regression equation. In other words, his policy is defined as our best prediction of what he will do (given our model), rather than what he actually does.

An index of policy agreement, or matching, must thus be based upon predicted scores. Such a matching index is obtained using the correlation between the estimates or predicted scores of the judges as follows:

$$\rho_m = \frac{E(Y_A' Y_B')}{\sigma_{Y_A'} \sigma_{Y_B'}} = \text{index of policy "matching", i.e., the extent to which the judges' policies are similar.} \tag{3}$$

If we say that a judge's policy is represented by the regression weights of his prediction equation (e.g., $Y'_A = \Sigma\, \beta_{Ai} X_i$), then any measurement of policy matching between two judges should measure the similarities between their respective regression weight vectors. One way to do this is to intercorrelate the two regression weight vectors or, as a substitute, intercorrelate the cue validity vectors (PROF technique of Wherry and Naylor, 1966).[2] However, these methods are subject to possible criticism if the matrix of covariances between all of the cues is significantly different from an identity matrix. For if the cues are not orthogonal, the meaningfulness of the relative magnitudes of either the regression weights or the cue validities must be questioned. (Hoffmann, 1962; Ward, 1962).

On the other hand, ρ_m can be interpreted in a very straightforward manner as the correlation between the prediction equations or hyperplanes. Consequently, ρ_m is clearly a measure of policy similarity. Whenever two policies are identical (that is, when $\beta_{Ai} = \beta_{Bi}$ for all $i = 1, \ldots, k$), ρ_m will equal one. Any deviations from $\beta_{Ai} = \beta_{Bi}$, other than an exception to be discussed below, will make $\rho_m < 1$.

There is one exception to the above rule, and that is when Y'_A is some linear function of Y'_B, i.e., $Y'_A = bY'_B + a$. However, we know that

$$E(Y_A') = E(Y_B') = 0;$$

therefore,

$$a = 0.$$

And since we know that

$$E(Y_A'^2) = \rho_A^2 \quad \text{and} \quad E(Y_B'^2) = \rho_B^2,$$

then

$$b = \frac{\rho_A}{\rho_B}.$$

The resulting relation is

$$Y_A' = \frac{\rho_A}{\rho_B} Y_B'.$$

---

[2] For a discussion of the relative merits of these two procedures see Naylor and Wherry, 1964.

And, it will also be true that

$$\beta_{Ai} = \frac{\rho_A}{\rho_B}\,\beta_{Bi}.$$

Consequently, when $\rho_m = 1$, we know that either the policies are identical or they are in proportion by a known constant. Whether two policies, related by the constant, ρ_A/ρ_B, are really different, is a matter of choice. Certainly, the only difference is in the consistency with which the judges adhered to their respective policies.

Tucker (1964) gives us another useful interpretation of ρ_m (or as he has named it, G). If we consider Y'_A and Y'_B to be vectors with lengths ρ_A and ρ_B, respectively, ρ_m is then the cosine of the angle between Y'_A and Y'_B (see Figure 2).

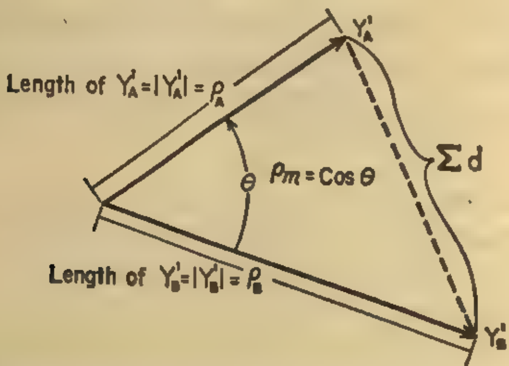


Figure 2. Pictorial representation of predicted score vectors, Y'_A and Y'_B, and their degree of similarity, ρ_m, suggested by Tucker, 1964.

It is interesting to note that Hursch, Hammond and Hursch (1964) did not mention ρ_m. Instead, they discussed a term, Σd, as the measure of agreement between a subject's policy and the policy (structure) of the ecology (which we have substituted with another subject or judge).

Using their symbols,

$$\Sigma d = \Sigma(\beta_{ei} - \beta_{si})(\rho_{ei} - \rho_{si}).$$

Substituting with the present notation, we have

$$\Sigma d = \Sigma(\beta_{Ai} - \beta_{Bi})(\rho_{Ai} - \rho_{Bi}).$$

This measure is clearly sensitive to any differences in policies and

can be shown to be directly related to ρ_m. In order to see this, consider the equation developed by Hursch, *et al.*, (their Equation 27, p. 55) to describe achievement or, in our case, response agreement:

$$\rho_a = \frac{R_e^2 + R_s^2 - \Sigma d}{2} + C\sqrt{(1 - R_e^2)(1 - R_s)^2}. \tag{4}$$

This can be rewritten in the notation used here as

$$\rho_a = \frac{\rho_A^2 + \rho_B^2 - \Sigma d}{2} + C\sqrt{(1 - \rho_A^2)(1 - \rho_B^2)}. \tag{5}$$

Another expression of ρ_a has been developed using ρ_m. From Naylor and Schenck (1965) we have

$$\begin{aligned}\rho_a &= E(Y_A Y_B) = E[(Y_A' + Z_A)(Y_B' + Z_B)]\\ &= E(Y_A' Y_B') + E(Y_A' Z_B) + E(Y_B' Z_A) + E(Z_A Z_B)\\ &= \rho_m \sigma_{Y_A'} \sigma_{Y_B'} + C\sigma_{Z_A}\sigma_{Z_B}\\ &= \rho_m \rho_A \rho_B + C\sqrt{(1 - \rho_A^2)(1 - \rho_B^2)}.\end{aligned} \tag{6}$$

The difference between Equations 5 and 6 is the first term on the right side of the equality, that is,

$$\frac{\rho_A^2 + \rho_B^2 - \Sigma d}{2} = \rho_m \rho_A \rho_B.$$

Therefore,

$$\Sigma d = \rho_A^2 + \rho_B^2 - 2\rho_m \rho_A \rho_B.$$

Thus, Hursch, *et al.*, have used a measure of policy matching which is simply the square of the distance between the two vectors Y'_A and Y'_B (see Figure 2).

Comparisons of ρ_a and ρ_m

If we consider Equation 6 and make the reasonable assumption that the residual scores for the judges are random, i.e., $E(Z_A Z_B) = 0$, then

$$\rho_a = \rho_m \rho_A \rho_B. \tag{7}$$

This relation states that response agreement is the degree to which the two judges successfully match their regression equations, weighted by the product of the intra-judge consistencies.

ρ_m is, as we have defined it, the correlation between the estimates

or predictions of the responses of judges A and B, that is, between Y'_A and Y'_B. Thus it is necessarily the similarity between the two prediction hyperplanes or better yet, between the two policies *unencumbered by error*. Consequently, ρ_m is an "error-free" index of policy similarity or agreement. It is different from ρ_a, response agreement, in that the latter is affected or attenuated by errors in prediction. This attenuation is measured by the indices, ρ_A and ρ_B. Equation 6 clearly shows that only when both policies are entirely predictable or captured, i.e., $\rho_A = \rho_B = 1.00$, will response agreement, ρ_a, be as high as policy agreement, ρ_m. Otherwise ρ_a will always be less than ρ_m.[8]

An interesting parallel can be drawn here with measurement theory. All practitioners in testing and measurement are familiar with the concept of "correction for attenuation." One can ask what the "true" relationship between two tests or variables would have been had there been no errors in measurement. The formula usually encountered is

$$\rho_{Y_1Y_2}(\text{true}) = \frac{\rho_{Y_1Y_2}}{\sqrt{\rho_{Y_1Y_1}\rho_{Y_2Y_2}}}.$$

In a like manner, we can ask what the agreement between two judges would have been had they been perfectly reliable. The important distinction here, however, is that unreliability is the failure to account for all of the judge's variation with a linear combination of the stimulus cues, i.e., the degree to which that judge is "inconsistent." With this in mind it can be seen that

$$\rho_A(\text{true}) = \rho_{Y_AY_A'}(\text{true}) = \frac{\rho_A}{\sqrt{\rho_{Y_AY_A}\rho_{Y_A'Y_A'}}}.$$

Now since ρ_A (true) represents the correlation between Y_A (true) and Y'_A (true), i.e., the correlation between a person's true score and his predicted true score, it must be unity. Similarly $\rho_{Y_A'Y_A'}$ must also be unity since it represents the reliability of a prediction equation based upon the entire population of elements. $\rho_{Y_AY_A}$ is *not* 1.00, however, since we still must consider error in the measurement of the judge's responses. Thus, we have

$$\rho_A = \sqrt{\rho_{Y_AY_A}}.$$

[8] This statement, of course, presumes that only positive or absolute values of correlations are being considered.

This equation bears out the difference between the two types of unreliability mentioned above, in that ρ^2_A, which is a measure of the degree to which we can predict that judge's responses from the stimulus values, is his reliability coefficient.

A similar derivation can be made for judge B so that we can now define "true" agreement as

$$\rho_a(\text{true}) = \frac{\rho_a}{\sqrt{\rho_{Y_A Y_A} \rho_{Y_B Y_B}}} = \frac{\rho_a}{\rho_A \rho_B} = \rho_m.$$

The result shows that if we approach ρ_m from an error of measurement point of view, it can be defined as the agreement between judges corrected for attenuation. On the other hand, ρ_m is also response agreement corrected for failure to capture, without error, the policies of the two judges.

In conclusion, the use of ρ_m as an index of rater agreement, either in place of or along with the more common index, ρ_a, would appear to offer interesting possibilities. As an index of policy agreement it is "error-free" and indeed can be viewed as ρ_a corrected for attenuation due to judge unreliability.

REFERENCES

Hammond, K. R., Hursch, C. J., and Todd, F. J. Analyzing the Components of Clinical Inference. *Psychological Review*, 1964, 71, 438-456.

Hoffman, P. J. Assessment of the Independent Contributions of Predictors. *Psychological Bulletin*, 1962, 59, 77-80.

Hursch, C. J., Hammond, K. R., and Hursch, J. L. Some Methodological Considerations in Multiple-Cue Probability Studies. *Psychological Review*, 1964, 71, 42-60.

Madden, J. M. *An Application to Job Evaluation of a Policy-Capturing Model for Analyzing Individuals and Group Judgment.* Lackland Air Force Base, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, May 1963. (PRL-TDH-63-15)

Naylor, J. C. and Schenck, E. A. A Revised Summary of the Multiple Cue Model and its Related Performance Indices. Statistical Parameters of Choice Behavior. Research Paper No. 2, June 1965. (Mimeo.)

Naylor, J. C. and Wherry, R. J. *Feasibility of Distinguishing Supervisors' Policies in Evaluation of Subordinates by Using Ratings of Simulated Job Incumbents.* Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, October 1964. (PRL-TR-64-25)

Naylor, J. C. and Wherry, R. J., Sr. The Use of Simulated Stimuli, Multiple Regression, and the JAN Technique to Capture and Cluster the Policies of Raters. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1965, 25, 969-986.

Peterson, C. R., Hammond, K. R., and Summers, D. A. Optimal Responding in Multiple-Cue Probability Learning. Behavior Research Laboratory Report No. 49, Institute of Behavioral Science, University of Colorado, 1965.

Schenck, E. A. and Naylor, J. C. Some Data Concerning Performance Indices Based upon the Multiple Regression Model when Applied in a Standard Multiple Cue Situation. Paper presented at Midwestern Psychological Association meetings, April 1965.

Smedslund, J. *Multiple Probability Learning*. Oslo: Akademisk Forlag, 1955.

Summers, S. A. The Learning of Responses to Multiple Weighted Cues. *Journal of Experimental Psychology*, 1962, 64, 29-34.

Tucker, L. R. A Suggested Formulation in the Development by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, 1964, 71, 528-530.

Uhl, C. N. Learning of Interval Concepts: I. Effects of Differences in Stimulus Weights. *Journal of Experimental Psychology*, 1963, 66, 264-273.

Ward, J. Comments on "the Paramorphic Representation of Clinical Judgment." *Psychological Bulletin*, 1962, 59, 74-76.

Wherry, R. J., Sr. and Naylor, J. C. Comparison of Two Approaches—JAN and PROF—for Capturing Rater Strategies. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 267-286.

Educational and Psychological Measurement
1966, 26, 825–831.

SIMILARITY ANALYSIS BY RECIPROCAL PAIRS FOR DISCRETE AND CONTINUOUS DATA

LOUIS L. McQUITTY
Michigan State University

The similarity index was developed as a central technique in an early method of hierarchical pattern analysis, called similarity analysis (McQuitty, 1955).

Similarity analysis has the advantage of being applicable to both discrete and continuous data. However, in its original form, it has the disadvantages of being complicated and laborious, sometimes requires iteration, and can lead to inconsistencies (McQuitty, 1955). By application of a particular theory of types (McQuitty, 1966) all of these problems can be solved, and a very simple method of hierarchical analysis can be developed, called Similarity Analysis by Reciprocal Pairs, applicable to both discrete and continuous data.

Theory

The theory says that every individual represents a succession of types, first an individual type, then types analogous to a species, a genus, a family, etc. As more and more individual types are classified together to represent higher and higher orders of hierarchical types, the successive categories become better representatives of pure types, which exist only in theory.

Individuals and hierarchical categories of the above kind are jointly characterized as typal representatives.

Every typal representative at any level x is best classified at the next higher level if it is classified with the typal representative most like it at level x and if the two representatives are reciprocal, i.e., Typal Representative i is most like j, and j is in turn most like i (McQuitty, 1957).

The Basic Equation

Let:

ij = any typal representative formed by combining the two typal representatives i and j from the next lower level, level x.

k = any typal representative other than i and j from level x.

a_{ik} = an index of association between i and k.

a_{jk} = an index of association between j and k.

a_{ij-k} = an index of association between ij and k.

Then:

$$a_{ij-k} = \frac{a_{ik} + a_{jk}}{2}. \tag{1}$$

An Illustration

In order to illustrate the method, it was applied to the data of Table 1, coefficients of correlation between people for the second

TABLE 1

Coefficients of Correlation between People

	B	C	D	E	F
B		60	49	24	43
C	60		62	46	51
D	49	62		40	57
E	24	46	40		67
F	43	51	57	56	

Note—Data from Stephenson (1953, p. 169); entries have been rounded from two to three places and decimal points have been omitted.

through the sixth persons of a matrix from Stephenson (1953, p. 169).

The coefficients of correlation were first converted to variances by squaring them and are shown in Matrix 1 of Table 2. This is not essential in most cases but does help increase the accuracy in details to the extent that variances represent better units than do coefficients of correlation.

The first step is to underline the highest entry in each column of Matrix 1, Table 2; it is 36 for Column B, 38 for C, 38 for D, and 31 for each E and F. Then, select the highest entry in the entire matrix; it is 38 and mediates between individuals C and D. In-

TABLE 2

A Similarity Analysis of Variance Indices Computed from the Correlation Coefficients of Table 1

Matrix 1

	B	C	D	E	F
B		36	27	06	18
C	$\underline{36}$		$\underline{38}$	21	26
D	27	$\underline{38}$		16	32
E	06	21	16		$\underline{31}$
F	18	26	32	$\underline{31}$	

Matrix 2

	B	CD	E	F
B		$\underline{32}$	06	18
CD	$\underline{32}$		19	29
E	06	19		$\underline{31}$
F	18	29	$\underline{31}$	

Matrix 3

	BCD	E	F
BCD		13	24
E	13		$\underline{31}$
F	$\underline{24}$	$\underline{31}$	

Matrix 4

	BCD	EF
BCD		19
EF	19	

dividuals C and D jointly are accepted as a better representative of a hierarchical type than is either separately. They are joined as a type and placed in both Row CD and Column CD of Matrix 2, Table 2.

Using the basic equation, the entry for both Row B—Column CD and Row CD—Column B is:

$$a_{CD-B} = \frac{a_{B-C} + a_{B-D}}{2} = \frac{36 + 27}{2} = \frac{63}{2} = 31.50, \text{ rounded to } 32. \quad (2)$$

The other entries of Row and Column CD are computed in an analogous fashion. The entries for the remaining cells of Matrix 2 are taken from the corresponding cells of Matrix 1.

The highest entry in each column of Matrix 2 is underlined and the highest entry in the matrix is chosen, 32 for Row B—Column CD and Column CD—Row B.

Typal Representatives B and CD are joined to form a higher Typal Representative, BCD, and are entered as a row and column of Matrix 3.

The basic equation is used to compute the entry for Row E—Column BCD and Row BCD—Column E.

$$a_{BCD-E} = \frac{a_{E-B} + a_{E-CD}}{2} = \frac{6 + 19}{2} = \frac{25}{2} = 12.50, \text{ rounded to } 13. \quad (3)$$

The entry for Row F—Column BCD and Row BCD—Column F is computed analogously. The other entries of Matrix 3 are taken from their corresponding cells of Matrix 2.

The highest entry in each column of Matrix 3 is underlined, and the highest entry in the entire matrix is determined, 31 for Row E—Column F and Row F—Column E. Typal Representatives E and F are therefore combined to form a new type at a higher level, Typal Representative EF. This completes the analysis. However, for purpose of further illustration, we show how the similarity index for Row BCD—Column EF and Row EF—Column BCD is computed:

$$a_{EF-BCD} = \frac{a_{E-BCD} + a_{F-BCD}}{2}$$

$$= \frac{13 + 24}{2} = \frac{37}{2} = 18.50, \text{ rounded to } 19. \quad (4)$$

A Critique

All reciprocal pairs of any matrix could have been operated on in preparing a new matrix. For example, Matrix 1 contains two reciprocal pairs, CD and EF. Each of these could have been combined to yield Matrix 5 as shown in Table 3.

TABLE 3

An Alternative Similarity Analysis of Matrix 1, Table 2

	B	CD	EF		BCD	EF
B		<u>32</u>	12	BCD		18
CD	<u>32</u>		24	EF	18	
EF	12	24				
	Matrix 5				Matrix 6	

The entry for Row B—Column CD and Row CD—Column B is computed in the same fashion as it was for Matrix 2, Table 2, and likewise for the entry of Row B—Column EF and Column EF—Row B.

The entry of Row CD—Column EF and Row EF—Column CD was, however, computed in a slightly different fashion; two applications of the basic equation were involved:

$$a_{CD-E} = \frac{21 + 16}{2} = 18.50, \text{ rounded to } 19. \tag{5}$$

$$a_{CD-F} = \frac{26 + 32}{2} = 29. \tag{6}$$

$$a_{CD-EF} = \frac{21 + 16 + 26 + 32}{4} = 23.75, \text{ rounded to } 24 \tag{7}$$

or

$$a_{CD-EF} = \frac{19 + 29}{2} = 24. \tag{8}$$

Stated in general notation, Equation 7 states,

$$a_{ij-kl} = \frac{a_{ik} + a_{il} + a_{jk} + a_{jl}}{2}. \tag{9}$$

This entire equation can be generalized to:

$$s_{xy} = \frac{\sum_{x=1}^{x=n} \sum_{y=1}^{y=m} a_{xy}}{n \cdot m}. \tag{10}$$

s_{xy} = the original similarity index (McQuitty, 1955).
x = Type x represented by Individuals $x_1, x_2, x_3, \cdots n$.
y = Type y represented by Individuals $y_1, y_2, y_3, \cdots m$.

The difference in Equations 1 and 10 derives from a difference in theory. Equation 1 says that each type should be given equal weight when they are joined to form an hierarchical type of the next higher level. Equation 10, on the other hand, says that the two types being joined should be given differential weights in relation to the number of individuals in each of them. Equation 7 (a special case of Equation 10) gave the same results as Equation 1, because in the special case of Equation 7 each CD and EF were composed of two individuals and were therefore given equal weights.

The theory of this paper recommends Equation 1; Equation 1 was derived from the theory of this paper and has the advantage of being very simple to apply to both discrete and continuous data.

An alternative theory of types could require differential weights in relation to the number of individuals in each of two types being joined to form a new type at the next higher level. An equation simpler to apply than Equation 10 can be developed for the purpose.

Let:

i = a type just realized by combining Types j and k; it will replace Types j and k in the next matrix.
j = a type composed of Individuals $b_1, b_2, b_3, \cdots b_n$.
k = a type composed of Individuals $c_1, c_2, c_3, \cdots c_m$.
l = any other type realized earlier by combining Types x and y.
x = a type composed of Individuals $d_1, d_2, d_3, \cdots d_r$.
y = a type composed of Individuals $e_1, e_2, e_3, \cdots e_s$.
a_{il} = the similarity index between Types i and l.
a_{jl} = the similarity index between Types j and l.
a_{kl} = the similarity index between Types k and l.

Then:

$$a_{il} = \frac{n(a_{jl}) + m(a_{kl})}{n + m} \tag{11}$$

Differential weights are used for Type i in the above application of Equation 11; they are represented by n and m, showing the numbers of individuals in each Type j and k which combined to yield Type i. Differential weights are not, however, used for Type l in

the above application of Equation 11; they were used earlier in the analysis when Equation 11 was applied to compute each a_{ji} and a_{kl}; they are represented by r and s.

Similarity Analysis by Reciprocal Pairs can be used to analyze any matrix (even one generated out of chance) into statistical types, for every matrix has at least one reciprocal pair. Consequently, the investigator should be especially aware of the need to find other evidence before interpreting the results as indicative of substantive types.

Summary

This paper generates an especially simple method for analyzing both continuous and discrete data into hierarchical types.

REFERENCES

McQuitty, L. L. A Method of Pattern Analysis for Isolating Typological and Dimensional Constructs, Lackland Air Force Base, Texas. *AFPTRC Research Bulletin*, 1955, *TN-55-62*.

McQuitty, L. L. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1957, 17, 207-229.

McQuitty, L. L. Single and Multiple Hierarchical Classification by Reciprocal Pairs and Rank Order Types. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 253-265.

Stephenson, W. *The Study of Behavior*. Chicago: The University of Chicago Press, 1953.

Educational and Psychological Measurement
1966, 26, 833–846.

THE RELATIVE EFFICIENCY OF TEST SELECTION METHODS IN CROSS-VALIDATION ON GENERATED DATA

LESTER C. SHINE II
The Ohio State University

The basic problem in test selection is one of selecting a subset of predictors (tests), given a larger set of possible predictors, in such a manner that a measured criterion can be predicted as efficiently or more efficiently (due to high shrinkage), in the probability sense, from a weighted composite of the subset of predictors as it can be from the whole set of weighted predictors. In other words, the aim of test selection is to maximize prediction of a given criterion and minimize cost where cost is a factor. To accomplish this aim, two common approaches have been followed. These approaches are statistical inference of predictor weights based upon a single sample (Wherry, 1951) and the technique of cross-validation (Mosier, 1951). The cross-validation approach has been studied here because it is the more realistic in the practical sense and because it overlaps with the statistical inference approach.

Implicit in the problem of test selection is the question of differential weighting of predictors, because no test selection can occur without differential weighting. Optimal weighting effectiveness usually occurs when there is considerable variation among the weights of a chosen weighting system (test selection method), when predictor intercorrelations are all small, and when correlations of the chosen weights with corresponding weights of other weighting systems are small (Scates and Noffsinger, 1931). It can be shown, however, that, under certain conditions, if the number of predictors is very large, the weighted composite predicts a given criterion equally well regardless of how the predictor weights are chosen (Wilks, 1938). Further, the presence of suppressor variables in a battery of

predictors may cause ordinary test selection methods to break down during the selection process itself (Wherry, 1946). Thus, it seems that, in addition to the obvious variable of sample size, the variables of number of potential predictors and homogeneity of potential predictors (as measured by size of test intercorrelations) are the more important variables to consider in an investigation of the relative efficiency of test selection methods in cross-validation.

The most comprehensive investigation to date of the relative efficiency of test selection methods in cross-validation has been conducted by Perloff (1951). Perloff manipulated sample size in his work while holding constant the variables of number of tests in the battery (seven) and homogeneity of the tests (high). His results indicated that the Standard Score Unit Weighting method worked best for small samples, Regression Analysis was best for very large samples, and the Wherry Test Selection method worked best for medium to large samples. Other less comprehensive comparative studies of test selection methods have been done by Anderson and Fruchter (1960), Grimsley (1949), Lawshe and Patinka (1958), and Lubin and Summerfield (1951). The gist of previous research indicated that the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus one or zero weights only, and the Wherry-Gaylord Test Selection method allowing plus integer or zero weights only were the best four methods to study under the manipulation of the variables of sample size, size of test battery, and homogeneity of test battery. Original sources for the Wherry Test Selection method and the two forms of the Wherry-Gaylord Test Selection method can be found in Stead, Shartle, *et al.* (1940) and Wherry and Gaylord (1946), respectively.

In general, previous studies on comparisons of test selection methods have been based on real data rather than computer generated data. Using computer generated data appeared to be better for the present investigation for two reasons. One reason was that the characteristics of computer generated data can be precisely controlled, whereas real data cannot usually be controlled so precisely. The second reason for using computer generated data instead of real data was that the size of the present investigation virtually precluded the possibility of using real data.[1]

[1] This article is based upon the author's dissertation done under Dr. Robert J. Wherry, The Ohio State University, 1965.

Method

Subjects

The subjects used in this experiment were entirely fictitious. The computer generated data used were equivalent to 40,000 subjects.

Apparatus

The only apparatus used in this experiment was the computer facilities of The Ohio State University (OSU) Computer Center, consisting of an IBM 1620 Computer and an IBM 7094 Computer.

Procedure

A fixed factor, four-way Analysis of Variance with no repeated measures and a within cell n of 5 seemed to be an appropriate analysis to use in the present investigation. The four factors were Sample Size, Size of Test Battery, Homogeneity of Test Battery, and Method, denoted by N, S, H, and M, respectively. The levels of N were chosen to be 50 and 200. The levels of S were chosen to be 6 and 12 (did *not* include the criterion). The levels of H were chosen to be Low and High, as measured by test intercorrelations. The levels of M were the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus one or zero weights only, and the Wherry-Gaylord Test Selection method allowing plus integer or zero weights only, denoted by WT, RA, $WG + 1$, and $WG + I$, respectively.

The basic data for the present study were randomly generated intercorrelation matrices having built in N, S, and H characteristics. These characteristics were built in through the use of the 1620 Score Generator Program available at OSU. This program made use of a theorem in Factor Analysis which says that $Z = \text{AF}$ (Harman, 1960), where Z is the matrix of observed scores on each test, A is the population factor structure composed of factor loadings for each test, and F is the matrix of scores on each factor variable, with all factors assumed to be mutually independent of each other. The program randomly drew samples of pre-assigned size N of factor scores from normally distributed independent populations, and then performed the matrix multiplication A times F to produce the matrix Z of observed test scores for each test. For data generation

purposes the criterion variable was considered to be the last test in the matrix Z. The tests were then intercorrelated to produce the observed intercorrelation matrix of predictors and criterion.

Since the matrix A entirely determined the characteristics of S and H, and since S and H each had two levels, four population matrices A were required for data generation. The number of rows of A controlled S, with 7 and 13 being used (the last row was reserved for the criterion). The pattern of loadings of A controlled H. Restrictions were imposed that the true population multiple R^2 between predictors and criterion range from .501 to .507 across the four A matrices and that the number of *common* factors be held constant at four for each of the four A matrices, with *unique* factors being included for each test. The four A matrices are presented in Table 1.

The proposed $2 \times 2 \times 2 \times 4$ ANOVA design with a within cell n of 5 implied that a total of 320 appropriate intercorrelation matrices were required. After the necessary 320 matrices were generated, the four test selection methods under consideration were applied, according to the above ANOVA design, to the first 160 matrices, hereinafter called validation matrices, of the 320 matrices to obtain test selection weights. This phase was called the validation phase of the experiment and its sole purpose was to obtain test weights from each of the 160 validation intercorrelation matrices. Standard criteria for stopping selection of tests were used where applicable. The 160 sets of test weights obtained in validation were then applied to the remaining 160 intercorrelation matrices, hereinafter called cross-validation matrices, using the standard formula for a composite correlation coefficient computed between a criterion and a set of predictors, to obtain cross-validation composite correlation coefficients. This phase of the experiment was called the cross-validation phase. The data obtained in this phase formed the basis for the ensuing data analysis.[2]

Results

The standard Fisher Transformation (Cramér, 1946), $Z_F|B = (1/2)\ ln\ (1 + R_c|B)/(1 - R_c|B)$, was applied to the composite

[2] Since this experiment may be replicated by using the four A matrices together with a score generator program, the generated score data are not available for distribution.

TABLE 1

The Four Factor Structures (A Matrices)[a]

	Low Homogeneity Factor					High Homogeneity Factor			
Test	1	2	3	4	Test	1	2	3	4
1	.30	.20			1	.80	.10		
2	.30	.60			2	.90	.30		
3	.20		.60		3	.90		.40	
4	.20		.20		4	.70		.10	
5	.40			.30	5	.80			-0-
6	.40			.10	6	.70			.20
C[b]	.30	.40	.50	.20	C[b]	.30	.40	.50	.20

	Factor					Factor			
Test	1	2	3	4	Test	1	2	3	4
1	.20	.10			1	.80	.20		
2	.10	.10			2	.80	.20		
3	.10	.80			3	.80	.30		
4	.20	.10			4	.80	.40		
5	.30		.10		5	.80		.10	
6	.30		.40		6	.80		.40	
7	.20		.30		7	.80		.20	
8	.20		.10		8	.80		.30	
9	.10			.10	9	.80			-0-
10	.30			.30	10	.80			.20
11	.30			.10	11	.80			.40
12	.10			.10	12	.80			.55
C[b]	.30	.40	.50	.20	C[b]	.30	.40	.50	.20

[a] The unique factor coefficients have been omitted. They may be calculated for any test, i, from the following formula:

$$u_i = (1 - \Sigma_j a_{ij}^2)^{1/2}$$

[b] C refers to the criterion.

correlation data to produce data suitable for analysis which transformed data is presented in Table 2 with decimal points omitted (decimal points appeared in front of each three digit number). A standard fixed-factor, four way ANOVA was then performed on the data presented in Table 2. The results of the ANOVA are presented in Table 3.

In order to allow for any possible biases in the F tests due to unequal cell variances (McNemar, 1949), a significance point of .025 was set, thereby insuring a true significance point not exceeding .050 in value. The only significant sources of variation were the main effect of N and the triple order interaction of $N \times S \times M$. The Newman-Keuls procedure for individual comparisons (Winer,

TABLE 2

Cross-validation Data After Fisher Transformation

H	S	N	Within Cell	M: WT	RA	WG + 1	WG + I
LO	6	50	1	250	509	385	389
			2	729	233	589	389
			3	237	422	900	333
			4	211	710	203	587
			5	504	460	539	090
			ΣX	1931	2334	2616	1788
			$\bar{X}$	386	467	523	358
		200	1	640	564	390	583
			2	624	469	569	533
			3	450	391	556	559
			4	411	585	295	560
			5	478	568	461	439
			ΣX	2603	2577	2271	2674
			$\bar{X}$	521	515	454	535
	12	50	1	328	482	480	534
			2	497	653	102	296
			3	395	458	358	455
			4	217	333	147	503
			5	459	544	−001	506
			ΣX	1896	2470	1086	2294
			$\bar{X}$	379	494	217	459
		200	1	555	527	377	413
			2	556	528	552	569
			3	665	451	591	462
			4	493	484	307	477
			5	348	427	496	461
			ΣX	2617	2417	2323	2382
			$\bar{X}$	523	483	465	476

1962) was applied to the $N \times S \times M$ effect table. The resulting ordered sums for the $N \times S \times M$ table are presented in Table 4. An underline joining any two sums in Table 4 indicates that the corresponding means (one-tenth of the sums) did not differ significantly at the .025 point. To aid in interpreting the results, a graphical presentation of the ordered means of Table 4 is presented in Figure 1.

TABLE 2 (Continued)

H	S	N	Within Cell	M: WT	RA	WG + 1	WG + I
			1	437	412	562	438
			2	377	665	504	692
			3	167	769	130	452
		50	4	400	484	509	353
			5	431	463	720	647
			ΣX	1812	2793	2425	2582
			$\bar{X}$	362	559	485	516
	6		1	482	631	509	458
			2	596	531	337	494
			3	598	499	459	466
		200	4	713	514	519	561
			5	524	417	455	469
			ΣX	2913	2592	2279	2448
			$\bar{X}$	583	518	456	490
HI			1	658	360	357	579
			2	400	400	349	697
			3	219	371	568	417
		50	4	320	372	525	474
			5	468	193	507	833
			ΣX	2065	1696	2306	3000
			$\bar{X}$	413	339	461	600
	12		1	444	334	469	481
			2	491	538	389	505
			3	383	419	429	409
		200	4	472	516	528	443
			5	400	497	520	528
			ΣX	2190	2304	2335	2366
			$\bar{X}$	438	461	467	473

Discussion and Conclusions

There is no need to explain or interpret the significant main effect of N since, according to previous research, this was to be expected. The significant interaction effect of $N \times S \times M$ indicated that sample size, size of test battery, and method jointly interacted among themselves. This joint effect can be more clearly grasped by considering what method produced the largest mean at each of the four possible treatment combinations of the levels of N and S. The

TABLE 3

ANOVA Source Table

Source	*SS*	*DF*	*MS*	*F*
N	109,093	1	109,093	7.24[a]
S	52,237	1	52,237	3.47
H	20,862	1	20,862	1.38
M	61,505	3	20,502	1.36
$N \times S$	1,012	1	1,012	<1[b]
$N \times H$	46,596	1	46,596	3.09
$N \times M$	87,373	3	29,124	1.93
$S \times H$	466	1	466	<1[b]
$S \times M$	70,352	3	23,451	1.56
$H \times M$	50,589	3	16,863	1.12
$N \times S \times H$	5,618	1	5,618	<1[b]
$N \times S \times M$	146,347	3	48,782	3.24[a]
$N \times H \times M$	56,594	3	18,865	1.25
$S \times H \times M$	101,309	3	33,770	2.24
$N \times S \times H \times M$	102,885	3	34,295	2.28
Error	1,929,174	128	15,072	—
Total	2,842,012	159	—	—

[a] Significant at the .025 point.
[b] Not significantly less than one at .025 point.

four possible combinations of N and S were 50(12), 50(6), 200(12), and 200(6), and the best methods for these combinations were $WG + I$, RA, WT, and WT, respectively. The unexpected result of RA working well for a small sample size and a small number of variables was difficult to explain except on the basis of sampling error. The mean at this treatment combination for $WG+1$ was very little different from that of RA as can be seen in Figure 1. This tended to support the sampling error explanation. The two means were not significantly different from each other according to Table 4.

The order of the four treatment combinations as laid out above appeared to be in the same order (descending) of expected overfit, holding method constant. Except for the questionable result for RA, the ordering of best methods corresponding to the treatment combinations as laid out above was about in ascending order of expected overfit, in regard to method alone. This result was reasonable and indicated that the variable of expected overfit was an important variable in determining what method worked best for a given combination of sample size and size of test battery. Another important variable seemed to be expected shrinkage which is very closely

TABLE 4

Ordered Means from $N \times S \times M$ Effect Table[a,b]

Treat. Comb.	50 (12) *WG*+1	50 (6) *WT*	50 (12) *WT*	50 (12) *RA*	50 (6) *WG*+*I*	200 (6) *WG*+1	200 (12) *WG*+1	200 (12) *RA*	200 (12) *WG*+*I*	200 (12) *WT*	50 (6) *WG*+1	200 (6) *WG*+*I*	50 (6) *RA*	200 (6) *RA*	50 (12) *WG*+*I*	200 (6) *WT*
Mean	339	374	396	417	437	455	466	472	475	481	504	512	513	517	529	552

[a] Any two means jointly underlined are *not* significantly different at .025 point.
[b] Decimals, originally appearing at the beginning of each three digit number, have been omitted.

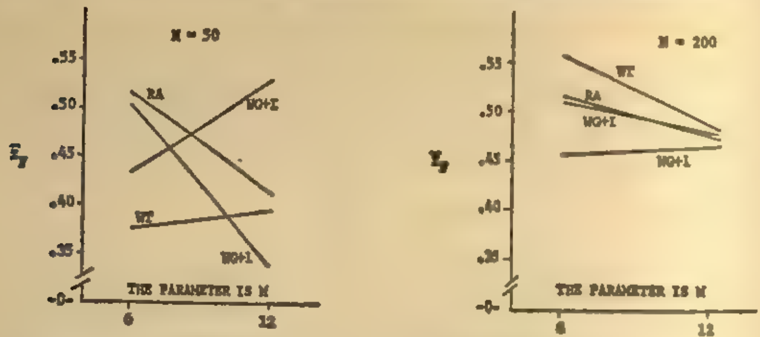


Figure 1.—$N \times S \times M$ Interaction

related to, but not completely determined by, expected overfit (overfit involves only one source of sampling error whereas shrinkage involves two sources of sampling error). It is probable that the operation of these two variables was the chief determiner of what method worked best for a given treatment combination. It should be noted that the ranking of methods in regard to expected shrinkage is, in general, the same as that for expected overfit.

The fact that no significant interaction effects other than $N \times S \times M$ were found in the present study suggested that an extreme sample size alone or an extreme size of test battery alone probably did not produce great differences among methods, or, in other words, that, for this case, any one of the four test selection methods probably would have worked equally well. Since H did not appear in any significant effect, it must be concluded that the size of predictor incorrelations was probably not an important variable in determining what method to use in a given situation. This is an important practical conclusion since it indicates that it is not necessary to evaluate predictor intercorrelations in determining what test selection method to use in a practical situation.

If the additional variable of cost or effort of method were taken into account, considerable additional information can be gleaned from the significant $N \times S \times M$ effect. The order (ascending) of the four methods under consideration according to cost or effort was $WG+1$, $WG+I$, WT, and RA, respectively. A consultation of Table 4 shows that the least costly methods, not significantly different from the best methods for the 50(12), 50(6), 200(12), and 200(6) treatment combinations, were $WG+I$, $WG+1$, $WG+1$, and $WG+I$, respectively. Thus it would seem that the Wherry-Gaylord Test Selection Method is a far better method than it has been given

credit for in the literature. This method also has the distinct advantages that it is equally applicable to test selection or item selection, and that it is easy to understand and apply. Further, the smallest and largest transformed means for the best methods and best least costly methods were .466 and .552, respectively. These values are equivalent to composite correlations of .435 and .502, respectively. In practice, such correlation values would frequently not be considered to differ enough to warrant extra expense.

The interpretation of the significant $N \times S \times M$ interaction has depended heavily upon the assumption that the assumed linear nature of the interaction would still hold if medium-valued levels had been included for N, S, and H. This assumption may not be warranted and should be checked in future research. If H were assumed to appear in no significant effects, the assumption could be investigated by allowing N and S to have three levels, say 6, 9, 12 and 50, 100, 200, respectively, with H being held constant at a medium level. If a within cell n of 5 were used along with the four methods studied in the present investigation, a total of 42,000 fictitious subjects would be required. Thus the size of this proposed experiment would be about the same as that of the present study.

No negative correlations were allowed in the population intercorrelation matrices used in this investigation. The effects of such correlations should be investigated in future research. If the assumption were made that H would not be involved in any significant effects, than a suitable investigation could be made by modifying the present design such that H is held constant at a medium level and that two additional method levels are added to M. The two new methods would be the Wherry-Gaylord Test Selection method allowing plus or minus unit weights or zero and the Wherry-Gaylord Test Selection method allowing plus or minus integer weights or zero. The population intercorrelation matrices should, of course, have both negative and positive correlations of such a nature that reversing test scoring would not remove the differential sign pattern. If a within cell n of 5 were used, a total of 30,000 fictitious subjects would be required for the proposed design.

Summary

The present investigation dealt with the relative efficiency of four representative test selection methods in cross-validation, using computer generated data. The four methods were chosen on the basis of

previous research and were the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus unit weights or zero only, and the Wherry-Gaylord Test Selection method allowing plus integer weights or zero only, denoted by *WT*, *RA*, *WG*+1, and *WG*+*I*, respectively. The efficiency of the four methods was studied under manipulation of the variables of sample size, number of predictors in the test battery, and homogeneity of the test battery (measured by size of predictor intercorrelations), denoted by *N*, *S*, and *H*, respectively, with method being denoted by *M*. A fixed factor, four-way ANOVA with no repeated measures and a within cell *n* of 5 was proposed to analyze the data. The levels of *N*, *S*, and *H* were 50 and 200, 6 and 12, and Low and High, respectively. The levels of M were the four chosen test selection methods.

The four methods were applied in validation to randomly drawn intercorrelation matrices with built-in appropriate *N*, *S*, and *H* characteristics. The resulting test selection weights were then applied in cross-validation to randomly drawn appropriate intercorrelation matrices to obtain cross-validation composite correlation coefficients. These coefficients were transformed, by using the standard Fisher Transformation, to insure normality of data. The proposed 2×2×2×4 ANOVA was then applied to the transformed data, with a significance point of .025 being set in advance. The only significant effects found were the *N* main effect, which was expected, and the $N \times S \times M$ interaction.

The best methods for the four *N*, *S* treatment combinations of 50(12), 50(6), 200(12), and 200(6) were *WG*+*I*, *RA*, *WT*, and *WT*, respectively. The result for 50(6) was questionable on a sampling error basis, and it was noted that for this particular treatment combination, *WG*+1 was only slightly less efficient than *RA*. It was concluded that the joint operation of the effects of expected overfit and expected shrinkage probably determined what test selection method works best for a given treatment combination. The fact that *H* appeared in no significant effect indicated that size of test intercorrelations was not an important variable to take into account in making a decision as to what test selection method to use in a given practical situation.

If the additional variable of cost or effort of method is taken into account, the best methods for the 50(12), 50(6), 200(12), and

200(6) treatment combinations of N and S were $WG+I$, $WG+1$, $WG+1$, and $WG+I$, respectively. It was concluded that the Wherry-Gaylord Test Selection method, with its advantages of being equally applicable to test or item selection and of being easy to understand and apply, is a far better test selection method than it has been given credit for in the literature.

Since the interpretation of the significant $N \times S \times M$ interaction depended heavily upon the assumption that its assumed linear nature would not change if medium-valued levels had been included for N and S, an experiment was proposed to investigate the validity of the assumption. Further, since no negative correlations were allowed in the population intercorrelation matrices from which sample matrices were drawn, an experiment was proposed to investigate the effects of such correlations upon the efficiency of test selection methods.

REFERENCES

Anderson, H. E. and Fruchter, B. Some Multiple Correlation and Predictor Selection Methods. *Psychometrika*, 1960, 25, 59-76.

Cramér, H. *Mathematical Methods of Statistics*. Princeton: University Press, 1946.

Grimsley, G. A Comparative Study of the Wherry-Doolittle and a Multiple Cutting-Score Method. *Psychological Monograph*, 1949, 63, 1-24.

Herman, H. *Modern Factor Analysis*. Chicago: The University of Chicago Press, 1960.

Lawshe, D. and Patinka, P. An Empirical Comparison of Two Methods of Test Selection and Weighting. *Journal of Applied Psychology*, 1958, 42, 210-212.

Lubin, A. and Summerfield, A. A Square-Root Method of Selecting a Minimum Set of Variables in Multiple Regression: II. A Worked Example. *Psychometrika*, 1951, 16, 425-437.

McNemar, Q. *Psychological Statistics*. New York: John Wiley and Sons, Inc., 1949.

Mosier, C. I. Problems and Designs of Cross-Validation. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1951, 11, 5-11.

Perloff, R. Using Trend-Fitting Predictor Weights to Improve Cross-Validation. Unpublished dissertation, The Ohio State University, 1951.

Scates, D. E. and Noffsinger, F. R. Factors Which Determine the Effectiveness of Weighting. *Journal of Educational Research*, 1931, 24, 280-285.

Stead, W. and Shartle, C., *et al. Occupational Counseling Techniques*. New York: American Book Company, 1940.

Wherry, R. J. Test Selection and Suppressor Variables. *Psychometrika,* 1946, 11, 239-247.

Wherry, R. J. Comparison of Cross-Validation with Statistical Inference of Betas and Multiple R from a Single Sample. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1951, 11, 23-28.

Wherry, R. J. and Gaylord, R. H. Test Selection with Integral Gross Score Weights. *Psychometrika,* 1946, 11, 173-183.

Wilks, S. S. Weighting Systems for Linear Functions of Correlated Variables Where There Is No Dependent Variable. *Psychometrika,* 1938, 3, 23-40.

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill Book Company, Inc., 1962.

Educational and Psychological Measurement
1966, 26, 847–851.

A RESEARCH NOTE ON THE METHOD OF ERROR-CHOICE

WARREN S. BLUMENFELD[1]
Purdue University

The method of error-choice has been described by Remmers (1954, p. 239), and it has been applied by Bernberg (1951), Hammond (1948), Kubany (1953), and Wechsler (1950a; 1950b). The method involves providing the subject with an unstructured, somewhat ambiguous, situation (question) in which his choice of item alternatives reveals his attitude. Operationally, this entails providing alternatives equi-distant from the true answer; but *not* providing a correct answer. The direction of deviation from the non-appearing correct alternative indicates the respondent's attitude.

Purpose

The purpose of this investigation was to determine the extent of relationship between an error-choice item and a criterion of affect for an attitude object.

Procedure

Data Collection

The data for this investigation were collected through the operations of the *Purdue Opinion Panel* (Gage and Remmers, 1948). A questionnaire, the topic of which was attitudes toward physicians, was administered nationally to high school students and to one parent of each student; it contained three item types—biographic,

[1] This report was adapted from a section of a doctoral dissertation submitted in partial fulfillment of the requirements of the Ph.D. degree at Purdue University, 1963. The author is now a Research Psychologist with the National Merit Scholarship Corporation at Evanston, Illinois.

semantic differential, and attitudinal. The research samples upon which this report is based contained 1000 students and one parent of each of the students, i.e., 1000 students and 1000 adults. The student sample was representative of the national population of high school students; the parent sample was defined by the student sample (Blumenfeld, Franklin, and Remmers, 1963).

Criterion

From the responses to the semantic differential scales, a composite image criterion for students and a composite image criterion for parents was developed which reflected individual differences in affect for the attitude object, physician. The interrater reliability of the student criterion was .82; the interrater reliability of the parent criterion was .83. The criterion development is described in detail by Blumenfeld (1963).

Error-Choice Item

The error-choice item administered to both students and parents was "How many years are there between the time a physician *first enters college* and the time he *completes internship?* The typical interim time, 9 years, did not appear; item alternatives provided were 6, 8, 10, and 12 years. The rationale of the error-choice item would indicate that selection of 6 or 8 connotes negative affect and that the selection of 10 or 12 connotes positive affect. That is, if the respondent holds physicians (the attitude object) in high esteem, he will attribute more academic preparation to physicians than is objectively the case; the more the positive affect, the more the academic preparation attributed.

Data Analysis

To determine the relationship of the error-choice item with the criterion, the hypothesis of no difference between the mean criterion, scores associated with item alternative selection was tested by means of analysis of variance (Winer, 1962, p. 96). In addition, to determine the relationship of the error-choice item with other items in the questionnaire, the item was tested by being paired with several other items which were judged to be pertinent. The hypothesis of zero interaction between item response distributions was tested by means of chi square (Siegel, 1956, p. 104).

Results

The analyses of variance in the student and parent samples yielded F statistics of 2.00 (df = 3,996) and 1.50 (df = 3,990), respectively. Neither of these obtained statistics is significant at the .05 level. That is, in both samples, the differences between the mean criterion scores associated with the item alternative selection could have arisen by chance.

Tables 1 and 2 summarize the chi square tests of interaction between the error-choice item and other items judged to be pertinent for the student and parent samples, respectively. Relationships significant at or beyond the .05 level are identified.

Discussion and Conclusion

Inspection of those items significantly related to the error-choice item in the student sample indicated that the higher the grade level in school, the longer the physician's educational preparation was thought to be; boys ascribed less time than did girls; white students ascribed more time to academic preparation than did others; better students perceived the process as longer than did poorer students; students who said they would want to become a physician

TABLE 1

Summary of Results of Chi Square Tests of the Hypothesis of Zero Interaction between Student Perception of Physician's Education and Selected Pertinent Variables

Variable	df	Chi Square
Grade in school	6	17.57**
Sex of student	3	45.72**
Socio-economic status	6	5.52
Political preference of student	6	3.22
Color	3	14.40**
Curriculum	9	16.45
Grades	9	19.64**
Religion	9	10.32
Desire of student to become physician	9	33.87**
Physician in family	3	0.29
Ben Casey viewing	9	444.09**
Dr. Kildare viewing	9	836.48**
Number of visits to physician in past year	9	14.08
Parent perception of physician's education	9	5.83

* Significant at .05 level.
** Significant at .01 level.

TABLE 2

Summary of Results of Chi Square Tests of the Hypothesis of Zero Interaction between Parent Perception of Physician's Education and Selected Pertinent Variables

Variable	*df*	Chi Square
Family breadwinner	3	2.28
Sex of responding parent	3	6.63
Physician	3	0.38
Age of parent	12	7.71
Education of parent	12	35.18**
Occupation	12	17.22
Political preference of parent	6	23.33**
Number of family visits to physician in past year	12	15.87
Desire for child to become physician	9	7.96
Ben Casey viewing	9	24.74**
Dr. Kildare viewing	9	19.75*
Attitude toward physician's fees	6	8.87
Attitude toward Medicare	9	9.12

* Significant at .05 level.
** Significant at .01 level.

saw the educational process as being shorter than did the students who were not interested; and those who viewed TV medic dramas more frequently felt that the educational process was shorter than did the less avid TV medic viewers.

For those items significantly related to the error-choice item in the parent sample, inspection indicated that the more education a parent had, the longer was the perception of education for physicians; Democrats tended to under-estimate the time while Republicans tended to over-estimate; and, as with the students, the more steady TV medic viewers tended to under-estimate the academic preparation while the less steady viewers tended to over-estimate.

In conclusion, the data in this investigation indicated that the error-choice item was *not* related to the criterion of attitude object affect in either the student sample or the parent sample; however, the item was related to student grade in school, sex of student, color of student, grades attained, desire of the student to become a physician, student viewing of TV medics, education of parent, political preference of parent, and parent viewing of TV medics.

REFERENCES

Bernberg, R. E. The Direction of Perception Technique of Attitude Measurement. *International Journal of Opinion and Attitude Research,* 1951, 5, 397-406.

Blumenfeld, W. S., Franklin, R. D., and Remmers, H. H. The Attitudes of Youth and Their Parents Toward Physicians. *Purdue Opinion Panel*, 1963, 22, Report No. 68.

Blumenfeld, W. S. The Image of the Physician: A Cross-Sectional Study of Attitudes and Their Correlates. Unpublished doctoral dissertation, Purdue University, 1963.

Gage, N. L. and Remmers, H. H. Opinion Polling with Mark-Sensed Punch Cards. *Journal of Applied Psychology*, 1948, 32, 88-91.

Hammond, K. R. Measuring Attitudes by Error-Choice: An Indirect Method. *Journal of Abnormal and Social Psychology*, 1948, 43, 38-48.

Kubany, A. J. A Validation Study of the Error-Choice Technique Using Attitudes on National Health Insurance. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1953, 13, 157-163.

Remmers, H. H. *Introduction to Opinion and Attitude Measurement.* New York: Harper, 1954.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

Wechsler, I. R. An Investigation of Attitudes Toward Labor and Management by Means of the Error-Choice Method: I. *Journal of Social Psychology*, 1950, 32, 51-62. (a)

Wechsler, I. R. A Follow-Up Study on the Measurement of Attitudes toward Labor and Management by Means of the Error-Choice Method. *Journal of Social Psychology*, 1950, 32, 63-69. (b)

Winer, B. J. *Statistical Principles in Experimental Design.* New York: McGraw-Hill, 1962.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 853–860.

COMPARISON OF PERSONALITY AND ATTITUDE VARIABLES

ANDREW L. COMREY
University of California, Los Angeles

APPLICATION of the factored homogeneous item dimension approach (Comrey, 1961) in a series of investigations (Comrey, 1962, 1964, 1965; Comrey and Schlesinger, 1962) has led to the development of a system of factors intended to describe certain aspects of the human personality. The same type of approach applied in the area of political and social attitudes resulted in the definition of several substantially correlated primary level factors which in turn produced a second-order general factor of Radicalism-Conservatism (Comrey and Newmeyer, 1965). The present investigation was undertaken for the purpose of determining the interrelationship, if any, of these two domains and in particular to ascertain whether any of the attitude measures might prove useful in measuring the main personality factors.

A 216 item inventory was prepared which consisted of 51 factored homogeneous item dimensions, or FHIDs (Comrey, 1961), a social desirability scale of eight items, and a four item validity scale designed to detect random marking. Each FHID consisted of four items which were written to measure a specifically conceived variable and hence to correlate substantially with each other. Each FHID, except one, was included in the analysis because it had a substantial loading on one of the personality or attitude factors in at least one previous analysis, or it was newly conceived for this investigation with the hope that it would be highly related to one of these factors.

Each FHID brought over from a previous study had emerged as an item factor in a factor analysis of items. Only items with load-

ings of approximately .45 or more were retained for inclusion within the FHID. Thus, items in a given FHID belong there by the joint criterion of logical and statistical affiliation. The personality factors of major importance were determined from correlations among FHIDs rather than from correlations among items. That is, each personality and attitude factor in this study was defined by highly loaded FHIDs rather than single items. This technique has the effect of providing a greater degree of stability to the structure of important factors than is normally likely to occur with analysis of single items. Furthermore, appearance of item specific factors tends to be restricted to the FHID level rather than creeping into the structure of the major factors. In using single items to search for the important factors, on the other hand, it is easy to produce factors which are more like FHIDs than important personality factors, since their most heavily loaded variables consist of very similar kinds of items.

Personality factors included in this study which were defined by four or more FHIDs were: Empathy, Neuroticism, Shyness, Compulsion, Dependence, and Hostility. Ascendance and Self Control were two additional personality factors defined respectively by three and two FHIDs each. Attitude factors included were: Welfare State Attitudes, Religious Attitudes, Punitive Attitudes, Nationalism, and Racial Tolerance Attitudes. Either two or three FHIDs were included to measure each of these attitude factors.

The items for a given FHID were well separated in the inventory booklet. A subject's response to each item was selected from one of two 9-choice scales and recorded on a separate answer sheet. The first response scale ranged from "1. Never" to "9. Always." The second ranged from "1. Absolutely Not" to "9. Absolutely." Volunteer university students and community persons ($N = 446$) served as subjects with a promise of an individual analysis of their test results as an inducement to participate. Community persons were obtained by requests for participation distributed in randomly selected blocks within a ten mile radius of UCLA. Subjects were about equally divided as to sex and the student-non-student status. The average age was between 26 and 27. The average education level was slightly more than one year of college. On a scale of political preference, the group showed a central tendency slightly on the democratic side.

A total score over each FHID for each subject was computed by adding up the numerical responses selected. These total scores were intercorrelated to provide a basis for dividing up the FHIDs into four item analysis groups. Highly correlated FHIDs were placed in different item analysis groups. Items from all the FHIDs included in a given item analysis group were then intercorrelated by Pearson r and factor analyzed for item analysis purposes. This was done to confirm previously developed FHIDs as item factors in this study and to check on new ones. If a FHID failed to emerge as an item factor in one of these analyses, it was dropped. Any item within a given FHID which failed to have an item factor loading of about .45 or more was dropped. Eight FHIDs lost one or two items out of four in this way. Correlations among FHIDs were corrected for attenuation in both variables to determine if any such value approached 1.00. This occurred for the FHIDs Racial Tolerance and Rapid Social Change, two FHIDs which defined an attitude factor in the previous study of attitudes. The five best items from these two FHIDs were combined into one FHID, named "Racial Tolerance." This was done to prevent the emergence of a factor in the analysis of FHIDs which would be merely a doublet based on these two alternate forms of the same FHID.

Total scores over the retained items in the 50 remaining refined FHIDs were calculated. These variables together with total scores over eight social desirability items and scores on four background variables provided the following list for the main analysis of FHIDs:

1. Acceptance of Authority, 2. Lack of Ego Strength, 3. Love of Routine, 4. Censorship, 5. Agreeableness, 6. Rhathymia, 7. Shyness, 8. Pacifism, 9. Status Quo, 10. Agitation, 11. Ascendance, 12. Racial Tolerance, 13. Cautiousness, 14. Service Orientation, 15. Reaction to Authority, 16. Weak Federal Government, 17. Need for Approval, 18. Talkativeness, 19. Depression, 20. Capital Punishment, 21. Defensiveness, 22. Order, 23. Cynicism, 24. World Government, 25. Self Sufficiency, 26. Desire for Power, 27. Social Desirability, 28. Religiosity, 29. Inferiority Feelings, 30. Need to Excel, 31. Personal Grooming, 32. Succorance, 33. Self Control, 34. Hostility, 35. Submission, 36. Friendliness, 37. Adequacy, 38. Conformity, 39. Welfarism, 40. Meticulousness, 41. Achievement, 42. Helpfulness, 43. Severe Treatment of Criminals, 44. Dependence,

45. Psychopathic Deviation, 46. Stage Fright, 47. National Service, 48. Pessimism, 49. Sympathy, 50. Drive to Finish, 51. Contraception, 52. Age, 53. Sex (1 male, 0 female), 54. Educational level, and 55. Political Preference (higher scores for liberals, lower scores for conservatives).

The 55 × 55 matrix of Pearson intercorrelations among these variables was factor analyzed by the minimum residual method, which requires no communality estimates since it operates on the off-diagonal elements only (Comrey and Ahumada, 1964). The minimum residual method automatically terminates extraction of factors when the largest remaining factor behaves as though it were associated with a negative eigen value. In this study, 19 factors were extracted before this occurred. Four of these factors had no loading greater than .3 and one additional factor had only one such loading, namely, .33. The factor solution was iterated with 14 factors, therefore, each time reinserting the communalities cumulated from the previous iteration. The communalities were reasonably stable after 10 iterations. The second and succeeding solutions were principal factor solutions, since the minimum residual method becomes a principal factor method when the diagonal cells are employed in the calculations.

The 14 stabilized unrotated factors were rotated analytically by the normal varimax method (Kaiser, 1958) and further rotated visually to oblique simple structure. The orthogonal solution is so close to the oblique solution, however, that only the orthogonal solution tables will be presented. Loadings of .3 or more on these 14 orthogonal rotated factors were as follows:

Empathy. Helpfulness, .76; Service Orientation, .69; Sympathy, .68; Social Desirability, .46; Dependence, .35; and Friendliness, .34.

Neuroticism. Depression, .78; Lack of Ego Strength, .73; Pessimism, .71; Agitation, .68; Adequacy, —.65; Inferiority Feelings, .60; Shyness, .36; Submission, .34; Hostility, .30.

Compulsion. Need for Order, .76; Drive to Finish, .62; Love of Routine, .62; Meticulousness, .56; Personal Grooming, .43; and Cautiousness, .41.

Shyness. Shyness, .76; Stage Fright, .58; Talkativeness, —.53; Desire for Power, —.40; Submission, .40; Friendliness ,—.40.

Hostility. Cynicism, .65; Hostility, .59; Psychopathic Deviation, .55; Rhathymia, .48; and Defensiveness, .39.

Dependence. Conformity, .70; Need for Approval, .68; Succorance, .46; Self Sufficiency, —.41; and Submission, .32.

Achievement Need. Achievement, .82; Need to Excel, .64; Desire for Power, .63; and Adequacy, .31.

Self Control. Self Control, .66; Agreeableness, .63.

Ascendance. Ascendance, .71; Submission, —.43; and Defensiveness, .33.

Welfare State Attitudes. Welfarism, .72; Weak Federal Government, —.72; Racial Tolerance, .66; World Government, .61; Political Preference, .55; Pacifism, .41; and Severe Treatment of Criminals, —.31.

Religious Attitudes. Censorship, .72; Contraception, —.70; Religiosity, .65; National Service, .32; Severe Treatment of Criminals, .31; and Rhathymia, —.30.

Punitive Attitudes. Capital Punishment, .54; Pacifism, —.50; National Service, .45; Severe Treatment of Criminals, .39; and Defensiveness, .35.

Sex. Personal Grooming, .51; Status Quo, .44; Educational Level, —.43; Sex, —.42; Reaction to Authority, —.42; National Service, .38; Social Desirability, .35; and Acceptance of Authority, .33.

Age. Age, .70.

Empathy, Neuroticism, Compulsion, Shyness, Hostility, and Dependence have been identified as major personality factors in several previous studies. Achievement Need emerged here for the first time in this series of investigations due to the addition in this study of the Achievement FHID to the list of variables. Self Control and Ascendance have appeared in previous studies but little success has been experienced to date in finding additional FHIDs to measure these factors.

Welfare State Attitudes, Religious Attitudes, and Punitive Attitudes each appeared as a factor in the previous study of political and social attitudes. The Racial Tolerance attitude factor from that study was not found here because the two FHIDs defining it were combined into one variable in the present study. The factor of Nationalism in that study, defined by National Service, Pacifism, and World Government failed to achieve separate definition

in the current analysis, but rather split up and combined with Welfare State Attitudes and Punitive Attitudes. There is a strong second order factor running through the attitude factors in both studies, however, and the present sample had far fewer politically conservative subjects. These facts may have been responsible for the disappearance of Nationalism as a separate factor here.

The major personality factors expected to emerge from this investigation did so in clear fashion. Only one of the major attitude primaries which could have been expected to emerge failed to do so. Two personality variables had small loadings on attitude factors, but no attitude variable had a loading as high as .3 on any personality factor. It would appear, therefore, that there is little hope of broadening the measurement base for these personality factors by using attitude variables of this kind.

Although an orthogonal solution was used, it is clear that correlations do exist among these factors, particularly the attitude factors. To study these interrelationships, factor scores were obtained for all factors except Age and Sex by adding item scores over those items belonging to FHIDs which had loadings of .39 or more on the factor. Exceptions were the omission of Social Desirability from Empathy, Submission from Shyness, and Pacifism from Welfare State Attitudes. Social Desirability was retained as a separate "factor" score while the other two FHIDs had higher loadings on other factors. In the case where a FHID had a negative loading on a factor, its item scores were reversed by subtraction from 10 before adding them into the factor score. Means, standard deviations, and split-half reliabilities corrected by the Spearman-Brown formula are given for these factor scores in the ADI materials.[1]

[1] The computations for this study were carried out on the IBM 7094 operated by the UCLA Computing Facility. The following materials have been deposited with the American Documentation Institute: matrix of correlations among the FHIDs and background variables (55 × 55); matrix of unrotated factors after ten iterative cycles; normal varimax rotated matrix; test booklet and answer sheet; list of FHIDs used in the factor analysis with their item numbers, means, standard deviations, and internal-consistency reliability estimates; list of factor scales, their means, standard deviations, reliability estimates, and the FHIDs defining them; matrix of correlations among the factor scores plus social desirability. Order Document No. 9039 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for 35 mm microfilm or $2.50 for photocopies. Make check payable to Chief, Photoduplication Service, Library of Congress.

The 12 factor scores and the Social Desirability scores were intercorrelated and factor analyzed by the minimum residual method. Only three factors could be extracted. These were rotated by the normal varimax method. The first had rotated loadings as follows: Punitive Attitudes, —.83; Welfare State Attitudes, .56; Religious Attitudes, —.52; and Compulsion, —.32. It is clearly the second-order factor of Radicalism-Conservatism also found in the first attitude study. The second factor had the following rotated loadings: Social Desirability, .77; Empathy, .70; Hostility, —.52; and Compulsion, .50. It can be referred to as "Social Desirability" without settling the question whether faking or sheer "goodness" is the main content of the factor. Probably some complex composite of the two is involved. The third second-order factor had the following loadings: Neuroticism, .74; Shyness, .56; Ascendance, .43; and Hostility, .31. This factor might be classified as Poor Social Adjustment.

The personality factors Empathy, Neuroticism, Hostility, Compulsion, Shyness, and Dependence have been repeatedly confirmed in this series of investigations. Welfare State Attitudes, Religious Attitudes, and Punitive Attitudes, and the general second-order factor of Radicalism-Conservatism defined by them, have been obtained twice. The scales measuring these particular factors, which are provided in the ADI materials, all have split-half reliabilities of .87 or more. Such scales should be useful in theoretical and applied research studies where personality or political and social attitude questionnaire measures are needed.

REFERENCES

Comrey, A. L. Factored Homogeneous Item Dimensions in Personality Research. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 417-431.

Comrey, A. L. A Study of Thirty-five Personality Dimensions. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 543-552.

Comrey, A. L. Personality Factors Compulsion, Dependence, Hostility, and Neuroticism. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 75-84.

Comrey, A. L. Scales for Measuring Compulsion, Hostility, Neuroticism, and Shyness. *Psychological Reports*, 1965, 16, 697-700.

Comrey, A. L. and Ahumada, A. An Improved Procedure and Program for Minimum Residual Factor Analysis. *Psychological Reports*, 1964, 15, 91-96.

Comrey, A. L. and Newmeyer, J. A. Measurement of Radicalism-Conservatism. *Journal of Social Psychology*, 1965, 67, 357-369.

Comrey, A. L. and Schlesinger, B. Verification and Extension of a System of Personality Dimensions. *Journal of Applied Psychology*, 1962, 46, 256-262.

Kaiser, H. F. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 1958, 23, 187-200.

Educational and Psychological Measurement
1966, 26, 861–870.

A FACTOR ANALYTIC STUDY OF THE SELF-RATINGS OF COLLEGE FRESHMEN

JAMES M. RICHARDS, JR.[1]
American College Testing Program

If a non-psychologist wished to learn something about another person, he would be very likely to ask a direct question, and he would frequently find out what he wanted to know. Psychologists, however, have neglected direct self-reports as a source of information about individuals, and have preferred test scores and ratings by teachers, supervisors, psychologists, etc. The reason for psychologists' infrequent use of direct self-reports is uncertain, but it is probably due to the ease with which self-reports can be distorted and because of a rather patronizing view of man's rationality, honesty, and self-insight.

There is, however, a growing body of evidence that suggests that the neglect of direct self-reports has unduly handicapped psychologists in their attempt to understand and predict behavior. For example, in one study, self-ratings were among the best predictors of the on-the-job performance of physical scientists (Taylor, Smith, Ghiselin, and Ellison, 1961), and in studies of high ability college students, self-ratings have been shown to contribute significantly to the prediction of high-level achievement in culturally important areas of behavior (Holland, 1963; Holland and Nichols, 1964). Similarly, Campbell (1959) has proposed that an important indicator of test validity should be that "test scores predict independent trait-appropriate or criterion measures better than do self-ratings" and pointed out that the available evidence (Campbell and Fiske, 1959) suggests that only rare tests can be considered valid by this standard.

[1] The author would like to thank John Holland, Sandra Lutz, and Clifford Abe for their many contributions to this project.

To use self-ratings effectively in research or practice, however, it is desirable to organize them into a relatively small number of categories. The present study, consequently, was conducted to develop a clear and sound basis for organizing self-ratings by college freshmen on a number of common traits into a brief profile. This brief profile can then be used in subsequent research to study more efficiently problems in which self-ratings provide useful information. The basic technique was to factor analyze 31 self-ratings for a broad cross-section of American college freshmen in diverse colleges. The present study parallels the author's earlier factor analysis of the life goals of college freshmen (Richards, 1965). It was believed that more useful and interpretable results would be obtained by separate factor analyses of life goals and self-ratings rather than by a single factor analysis of goals and ratings like that done by Astin and Nichols (1964).

Method

The present study grew out of the American College Survey (Abe, Holland, Lutz, and Richards, 1965) which was conducted by the American College Testing Program to obtain a more complete account of the typical American college student and the variation among students from college to college. To accomplish this task, a comprehensive assessment was administered in the months of April or May, 1964, to 12,432 college freshmen in 31 institutions of higher education including liberal arts colleges, state universities, teachers colleges, and two-year community colleges. This group of freshmen, of whom 6289 were men and 6143 were women, provided the sample for this study.

In this study sample, 7 per cent were enrolled in junior colleges, 12 per cent in four-year undergraduate colleges, and 81 per cent in universities offering graduate work. Approximately 15 per cent were students in private colleges, while 85 per cent were students in public colleges. About 95 per cent attended coeducational colleges. Finally, 20 per cent were enrolled in colleges in the Northeast, 31 per cent in colleges in the South, 20 per cent in colleges in the Midwest, 26 per cent in colleges in the Mountains and Plains states, but only 3 per cent in colleges on the West Coast. From these figures it would appear that students in coeducational colleges are somewhat over-represented and students in West Coast

colleges are considerably under-represented in the sample. Nevertheless, the over-all characteristics of the sample suggest that it represents a reasonable cross-section of American college freshmen in 1964.

Scores on a nationally administered test of academic potential (the ACT test battery) were available for 7262 of these freshmen. A comparison was made between the distribution of test scores in this sub-sample and the corresponding distribution in a national norm group (Holland and Richards, 1965). The results revealed that on all ACT subtests the sample includes fewer persons with low scores than does the national norm group. These differences probably occurred because the norm group consisted of potentially college-bound high school seniors while the sample consisted of college freshmen who had already survived more than one half of the academic year. However, a full range of talent is represented in the sample, and it differs only slightly from the national distribution.

The assessment device used to collect data was the American College Survey (Abe et al., 1965), a booklet which contains a letter explaining the purpose of the survey and a series of sections planned to elicit information about each student's achievement, aspirations, attitudes, interests, potentials, values, and background. The American College Survey was administered at each college by personnel at that college. The survey was filled out by students—who recorded their 1004 responses on two special answer sheets—in English classes, chapels, and convocations or in dormitories and their homes.

For the present study, thirty-one self-ratings on common traits were used. Each of the subjects rated himself on each of the thirty-one traits on a four-point scale ("Below Average," "Average," "Above Average," "Top Ten Per Cent"), and scores from one to four were assigned to these responses so that a higher score indicated a greater possession of the trait in question. A complete list of these traits, together with the means and standard deviations for each sex is shown in Table 1.

Product moment correlations between the self-rating items were computed separately for each sex.[2] The two resulting 31×31

[2] Calculations were carried out at the Measurement Research Center, University of Iowa and at the University of Utah Computer Center.

matrices were factor analyzed using the principal components method based on eigenvalues and eigenvectors with unity in the diagonal and extracting all factors with an eigenvalue greater than 1.00. This procedure, including the use of unity in the diagonal, is Harris's (1964) Model A factor analysis, and follows the rationale presented by Kaiser (1960). An orthogonal rotated solution was computed by the Varimax procedure (Kaiser, 1958), and an oblique rotated solution by the Promax procedure[3] (Hendrickson and White, 1964) with $k = 4$.

Results

For each sex, seven factors had an eigenvalue greater than 1.00 and were included in the rotations. For both sexes, the total .10 hyperplane count is substantially higher for the oblique solution than for the orthogonal solution. This indicates clearly that the oblique solutions should be used in the further interpretation of results. Accordingly, the oblique rotated factors for both sexes obtained by the Promax procedure (Hendrickson and White, 1964) are shown in Table 2, and the correlations among Promax factors in Table 3.

The *Coefficient of Congruence* (Tucker, 1951) was computed between each Promax factor for males and each Promax factor for females to determine the extent to which the structure of self-ratings is similar for men and women. These coefficients are presented in Table 4 with female factors rearranged to place highest *Coefficients of Congruence* in the diagonal.

Discussion

The results shown in Table 4 indicate a high degree of similarity between the structure of self-ratings for males and the corresponding structure for females, since a good match is obtained for all seven factors for both sexes. Moreover, the match has both

---

[3] Tables showing for each sex the correlations among the self-ratings, the Varimax orthogonal rotated factors, and the transformation matrices for computing the Promax solutions from the Varimax solutions have been deposited as Document number 9040 with the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. A copy may be secured by citing the Document number and by remitting $1.25 for photoprints, or $1.25 for 35 mm. microfilm. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

Means and Standard Deviations of Self-Ratings for Both Sexes

Self-Rating	Males Mean	Males S.D.	Females Mean	Females S.D.
1. Originality	2.44	.74	2.34	.70
2. Leadership	2.43	.79	2.28	.78
3. Mechanical Ability	2.24	.92	1.64	.75
4. Popularity	2.44	.70	2.35	.64
5. Athletic Ability	2.48	.89	2.15	.81
6. Understanding of Others	2.73	.73	2.82	.69
7. Drive to Achieve	2.62	.80	2.60	.77
8. Mathematical Ability	2.33	.94	1.97	.89
9. Scholarship	2.35	.78	2.39	.78
10. Sociability	2.41	.74	2.51	.71
11. Artistic Ability	1.74	.85	1.84	.85
12. Aggressiveness	2.12	.74	2.07	.70
13. Speaking Ability	2.23	.78	2.17	.73
14. Self-Control	2.61	.78	2.51	.70
15. Independence	2.81	.77	2.68	.76
16. Scientific Ability	2.14	.85	1.73	.75
17. Conservatism	2.26	.74	2.18	.65
18. Practical Mindedness	2.52	.71	2.51	.69
19. Writing Ability	2.13	.77	2.19	.73
20. Expressiveness	2.24	.71	2.26	.71
21. Cheerfulness	2.53	.72	2.60	.70
22. Self-Confidence (Social)	2.27	.81	2.17	.77
23. Self-Confidence (Intellectual)	2.34	.74	2.18	.70
24. Perseverance	2.37	.67	2.36	.64
25. Popularity with the opposite sex	2.34	.73	2.26	.70
26. Research Ability	2.11	.69	1.91	.64
27. Physical Energy	2.65	.76	2.38	.71
28. Sense of Humor	2.76	.72	2.66	.67
29. Physical Health	2.82	.81	2.66	.79
30. Acting Ability	1.88	.79	1.91	.78
31. Sensitivity to the needs of others	2.55	.73	2.73	.71

convergent and discriminant validity (Campbell and Fiske, 1959). The *Coefficients of Congruence* between matching factors range from .90 to .99 with a median of .94, while the coefficients for unmatched factors ranged from —.12 to .33 with a median of .01. It should be emphasized that the factor analyses and rotations were completely independent, and that the samples were large and diverse. These results, therefore, are impressive evidence for a consistent organization of self-ratings, or of self-concepts, for both sexes. Such consistency from sample to sample is an important consideration in determining the adequacy of representation of the domain by the rotated factor solution (Harman, 1960).

TABLE 2

Promax Oblique Rotated Factor Matrix for Each Sex

	Males							Females						
	A	B	C*	D	E	F*	G	A	B	C	D	E*	F*	G
1.	05	01	72	02	18	−07	04	02	02	−06	−02	62	23	13
2.	08	04	14	13	−02	45	−16	56	23	00	03	12	−01	−15
3.	05	13	−01	19	82	10	−13	06	−02	06	18	−03	74	−02
4.	08	05	−09	−10	−03	79	10	80	07	−04	−10	−12	02	12
5.	86	−01	−07	−17	−03	06	−05	07	−07	71	−05	−06	25	−12
6.	05	00	12	43	−17	00	50	04	05	−09	12	10	02	70
7.	05	32	−01	40	−19	02	−13	02	59	04	15	06	−13	05
8.	01	86	−26	−09	10	05	03	07	73	−06	−12	−31	31	05
9.	−01	75	07	−04	−20	−03	−03	−02	85	−08	−11	08	−04	03
10.	−13	01	00	−03	00	84	19	76	02	−10	−05	−06	−04	24
11.	−04	−15	73	−05	45	−09	−02	−05	−20	−10	−03	57	52	13
12.	13	−10	13	26	06	38	−32	61	−08	00	13	21	06	−28
13.	−08	−10	54	07	−07	26	−08	29	01	00	11	51	−10	−14
14.	−01	−04	−03	70	02	−07	34	02	00	−02	57	−05	02	26
15.	11	−08	13	63	15	−10	07	11	03	09	46	08	05	06
16.	−04	73	01	05	44	−06	02	−05	52	03	00	−05	51	−04
17.	−14	−01	−26	72	15	01	02	−04	−07	−08	79	−13	14	−12
18.	−07	00	−14	79	11	02	19	−01	−04	−01	70	−12	09	19
19.	00	19	75	−22	−24	−15	14	−18	20	−05	−08	77	−10	07
20.	−02	08	74	−11	−14	02	10	03	07	−03	−07	75	−09	07
21.	−04	07	−07	01	00	61	46	38	−02	19	00	−09	−06	39
22.	−08	−05	05	04	07	80	01	77	−03	01	07	02	−04	−07
23.	−06	53	21	05	−04	15	−03	15	59	02	02	21	−06	−06
24.	05	22	06	46	−07	−03	−02	−08	39	13	31	13	−10	03
25.	07	00	01	−10	09	76	08	75	−03	−03	−03	−01	08	05
26.	01	55	26	05	31	−08	−01	−09	36	11	06	23	34	−04
27.	88	00	−01	00	07	−04	05	−05	00	90	−03	02	02	−04
28.	21	06	13	−09	03	33	51	13	−07	33	−13	13	01	45
29.	81	−04	−03	06	03	−05	17	−12	02	79	04	−05	−12	08
30.	−07	−11	64	−09	09	16	12	11	−12	10	−11	59	11	03
31.	−01	−09	17	39	−14	−03	60	−10	02	−07	09	09	00	78
.10 hyper-plane count	23	20	14	17	17	19	17	17	20	24	17	15	20	16

* Reflected factor.

TABLE 3

Correlations among Promax Oblique Factors

	A	B	C	D	E	F	G
A	—	24	26	39	−11	51	03
B	27	—	39	46	−12	26	−09
C	44	21	—	52	−24	45	−04
D	31	39	30	—	−25	46	−13
E	46	33	22	33	—	−23	19
F	−03	09	08	−07	−01	—	−02
G	38	19	32	29	25	−01	—

Note—Correlations for males are shown above the diagonal and for females below. Factors are reflected as appropriate.

TABLE 4

Coefficients of Congruence between Male and Female Promax Factors

	Males						
Females	A	B	C[a]	D	E	F[a]	G
C	94	01	−04	−03	09	03	15
B	−01	93	05	12	−06	02	−04
E[a]	−02	01	99	−06	01	00	08
D	−06	−05	−12	94	22	−04	16
F[a]	08	33	16	11	90	−01	−06
A	00	−01	06	03	07	96	07
G	00	01	16	31	−12	14	93

[a] Reflected factor.

The Promax factors are briefly described and interpreted below:

Male A-Female C has high loadings for both sexes on the traits: "athletic ability," "physical energy," and "physical health" with a secondary loading on "sense of humor." An appropriate title would be *Physical Well Being*.

Male B-Female B has high loadings on the traits: "mathematical ability," "scholarship," "scientific ability," and "self-confidence (intellectual)," with secondary loadings on "drive to achieve" and "research ability." This pattern could be summed up by the title *Scholarship*.

Male C-Female E has high loadings on "originality," "artistic ability," "speaking ability," "writing ability," "expressiveness," and "acting ability." An obvious title would be *Estheticism*, and one would expect scores on this factor to be correlated with independent estimates of originality. This pattern of loadings implies that college freshmen conceive of creativity largely in terms

of the arts, and have little conception of the role of originality in the sciences or other non-artistic fields.

Male D-Female D has high loadings on "self-control," "independence," "conservatism," "practical mindedness," and "perseverence" for both sexes. In addition, it has moderate loadings for males, but not for females on the traits "understanding of others" and "drive to achieve." This cluster of traits suggests a well-integrated, self-reliant, and pragmatic approach to life. This factor thus appears to tap some of the same psychological characteristics as the "Achievement via Independence" Scale of the CPI (Gough, 1957). Perhaps an appropriate title for this factor would be *Pragmatism.*

Male E-Female F has high loadings on "mechanical ability," "artistic ability," "scientific ability," and "research ability." An appropriate title might be *Technical-Scientific Ability*. The high loading for "artistic ability" might indicate that this factor taps scientific creativity.

Male F-Female A has high loadings for both sexes on the traits "leadership," "popularity with opposite sex," "sociability," "aggressiveness" (with a substantially higher loading for females than for males), "cheerfulness," "self-confidence (social)," and "popularity." The overall impression created by these loadings is of a pleasant, confident extrovert. The most appropriate title might be *Sociability.*

Male G-Female G has high loadings on "understanding of others," "sensitivity to the needs of others," "cheerfulness," and "sense of humor," with secondary loadings on "aggressiveness," and "self-control." This configuration of traits is close to what is called "empathy," and an appropriate title might be *Sensitivity to Others.*

These self-rating factors, in combination with the factors obtained in the earlier study of life goals (Richards, 1965) may lead to better studies of vocational choice, vocational counseling, and the prediction of vocational success by making it possible to control efficiently for both present status and for intentions for the future. Such control is especially critical for studies of vocational success, since if initial self-ratings were not considered, the variables identified as predicting success might be merely the variables which correlated with original self-concepts. In such a case,

it would be better to use self-ratings directly rather than other variables which indirectly reflect the same characteristics.

Summary and Conclusions

The primary purpose of this study was to provide a brief profile of self-ratings which would adequately describe some common characteristics of the self-conceptions of American college freshmen. In samples of 6289 male and 6143 female college freshmen, 31 self-ratings on common traits were intercorreleted. With unity in the diagonal, a principal components analysis was carried out, extracting all factors with an eigenvalue greater than 1.00. An orthogonal rotated solution was computed by the Varimax procedure, and an oblique solution by the Promax procedure. The oblique solutions were more satisfactory, and were used in assessing the similarity of the two rotated matrices. Seven factors common to both sexes were titled: Physical Well Being, Scholarship, Estheticism, Pragmatism, Technical-Scientific Ability, Sociability, and Sensitivity to Others. No factors unique to one sex were obtained.

The study appears successful in attaining its main purpose, since the number of variables was reduced from 31 to 7 for each sex. There is a high degree of congruence between the factors for males and females; the obtained factors are easily interpreted; and the use of large diverse samples lends strong support to our confidence in the factor pattern. The reduction of self-ratings to seven factors provides a simple, brief set of items for assessing self-concepts in questionnaires and similar assessment devices so that more expensive and time-consuming devices are not always necessary or desirable.

REFERENCES

Abe, C., Holland, J. L., Lutz, Sandra W., and Richards, J. M., Jr. *A Description of American College Freshmen.* ACT Research Report No. 1. Iowa City: American College Testing Program, 1965.

Astin, A. W. and Nichols, R. C. Life Goals and Vocational Choice. *Journal of Applied Psychology*, 1964, 48, 50-58.

Campbell, D. T. Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity. *American Psychologist*, 1959, 25, 546-553.

Campbell, D. T. and Fiske, D. W. Convergent and Discriminant

Validation of the Multitrait-Multimethod Matrix. *Psychological Bulletin,* 1959, 56, 81-105.

Gough, H. G. *Manual for the California Psychological Inventory.* Palo Alto, California: Consulting Psychologists Press, 1957.

Harman, H. H. *Modern Factor Analysis.* Chicago: University of Chicago Press. 1960.

Harris, C. W. Four Models for Factor Analysis. Paper read at American Psychological Association, Los Angeles, 1964.

Hendrickson, A. E. and White, P. O. Promax: A Quick Method for Rotation to Oblique Simple Structure. *British Journal of Statistical Psychology,* 1964, 17, 65-70.

Holland, J. L. The Prediction of Achievement in Different College Environments. Paper read at American Psychological Association, Philadelphia, 1963.

Holland, J. L. and Nichols, R. C. Prediction of Academic and Extracurricular Achievement in College. *Journal of Educational Psychology,* 1964, 55, 55-65.

Holland, J. L. and Richards, J. M., Jr. *Academic and Nonacademic Accomplishment: Correlated or Uncorrelated?* ACT Research Report No. 2. Iowa City: American College Testing Program, 1965.

Kaiser, H. F. The Application of Electronic Computers to Factor Analysis. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 141-151.

Kaiser, H. F. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika,* 1958, 23, 187-200.

Richards, J. M., Jr. Life Goals of American College Freshmen. Paper read at American Psychological Association, Chicago, 1965.

Taylor, C. W., Smith, W. R., Ghiselin, B., and Ellison, R. S. *Explorations in the Measurement and Prediction of Contributions of One Sample of Scientists.* United States Air Force, Personnel Lab, Lackland Air Force Base, ASD-TR-61-96, 1961.

Tucker, L. R. A Method of Synthesis of Factor Analysis Studies. *Personnel Research Section Report,* No. 984, Washington, D. C.: Department of the Army, 1951

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 871–882.

JUDGING COMPLEX VALUE STIMULI: AN EXAMINATION AND REVISION OF MORRIS'S *PATHS OF LIFE*[1]

PAUL DEMPSEY
Institute of Behavior Assessment, Davis, California
AND
WILLIAM F. DUKES
University of California, Davis

DURING the past few decades the philosopher Charles Morris has sought to examine individuals' ethical and religious orientations using a quantitative approach. In his investigations he has developed an instrument which he has called "Paths of Life" or "Ways to Live." It consists of 13 paragraphs, each presenting a conception of "the good life." Morris (1956) has epitomized these paragraphs as follows:

Way 1: Preserve the best that man has attained.
Way 2: Cultivate independence of persons and things.
Way 3: Show sympathetic concern for others.
Way 4: Experience festivity and solitude in alternation.
Way 5: Act and enjoy life through group participation.
Way 6: Constantly master changing conditions.
Way 7: Integrate action, enjoyment, and contemplation.
Way 8: Live with wholesome, carefree enjoyment.
Way 9: Wait in quiet receptivity.
Way 10: Control the self stoically.
Way 11: Meditate on the inner life.
Way 12: Chance adventuresome deeds.
Way 13: Obey the cosmic purposes.

[1] Financial support for this research was provided by the University of California Committee on Research. Thanks are expressed to Charles Morris for his comments on a preliminary draft of the manuscript.

The responder's task in completing the instrument is first to rate each of the 13 descriptive paragraphs on a 7-point scale, and then to rank them according to his preferences. Using data obtained in this manner, Morris has explored differences in philosophies of life according to national groupings, religious affiliation, economic status, temperament, character, body-type, etc. His principal findings, together with a copy of the "Ways to Live" document, are contained in his book, *Varieties of Human Value* (Morris, 1956). Other reports of his investigations have appeared in psychological periodicals—e.g., Osgood, Ware, and Morris (1961) and Morris and Jones (1955).

Because Morris has been attempting to assess persons' *Weltanschauungen*, his work should be of interest to behavioral scientists in general and of special concern to personality theorists. But due both to the breadth of the general task he has undertaken and to some features of the particular instrument with which he has approached it, Morris is naturally a target for critical comment. One focus for complaint is Morris's use of language. His phrasing, often poetic, while understandable as an attempt to reflect with minimum distortion the tone of ethical and religious teachings, nonetheless lacks the rigor psychologists have come to expect in psychometric work. One finds, for example, these statements embedded in the longer paragraphs: *Life continuously tends to stagnate, to become comfortable, to become sicklied o'er with the pale cast of thought; To sit alone under the trees and the sky, open to nature's voices, calm and receptive, then can the wisdom from without come within; One should be a serene, confident, quiet vessel and instrument of the great dependable powers which move to their fulfillment*. The effect of such a mode of expression is difficult to assess. (See Winthrop, 1959, for an extensive critique.)

Problem

The present paper is less concerned with the manner of expression of ideas, however, than with what may be called the incongruity between stimulus complexity and response simplicity. For each of the 13 "supermolar" situations the responder is expected to weigh his overall feelings, synthesize his attitudes, and code these into a simple number which constitutes his response. Compared to

the usual personality, attitude, interest inventory, the scope of the input unit here (a paragraph) is unusually large while that of the output unit (a number) is standard. In the face of this discrepancy two focal questions are being raised: (1) Are there confounding elements in the paths which elicit partial responses discordant with the response trends to the paragraphs as wholes? (2) Can the paragraphs be simplified without destroying the holistic scope of the document? These may be called questions of homogeneity and simplicity; positive answers to them entail questions of reliability which will also be considered here.

Procedure

The study was conducted in three phases: a normative or preparatory phase, a cross-validation, and a check on reliability. A total of 230 *S*s, all members of undergraduate psychology classes at the University of California, Davis, were used.

Phase 1. (a) The 13 paths were broken down into 110 sentences, some obtained by taking, intact, short sentences from the original paragraphs and the remainder by making two or more simple statements from more complex ones. (b) 24 *S*s responded to the original document and performed normalized Q sorts of the 110 statements. (c) On the basis of within-path inter-item *r*'s calculated from the Q sorts, two other versions of the document were prepared: (1) a *revised* form from which empirically discordant elements had been eliminated, and (2) a *short* form containing only the most highly intercorrelating elements.

Phase 2. (a) 40 *S*s ranked and rated both the revised and original forms and also made Q sorts of the 110 statements. (b) The paragraphs in the original, revised, and short forms were compared for mean differences in within-path inter-item *r*'s (part-part relationships) and inter-path *r*'s (part-whole relationships).

Phase 3. (a) With appropriate counterbalancing 102 *S*s completed either two of the three forms or the short form twice, an interval of 10 days occurring between tasks. (b) 64 additional *S*s completed the original, revised, or short form only. (c) Mean test-retest and alternate-form reliabilities were estimated, using these 166 *S*s plus the 64 from Phases 1 and 2.

Results and Discussion

1. *The Question of Discordant Elements.*

On the assumption that even in a small sample, negative correlation coefficients can be taken as crude indicators of discordancy, an intercorrelation matrix for each path was derived from the information provided by the Q sort in Phase 1. That is, for Path 1, which had been separated into 10 statements, 45 r's were computed, the *Ss*' Q scores for statement 1 being correlated with their Q scores for statement 2, etc., and comparable part-part correlations were computed between all possible pairs of statements in each of the remaining 12 paths. Thus 13 matrices of r's were assembled.

In all but two of the matrices negative coefficients appeared; in some more than half the entries were negative. In Path 4, for example, the statements, *The aim of life should not be to control the course of the world or society or the lives of others* and *The aim of life should be to be open and receptive to things and persons, and to delight in them* correlated positively with each other, but both showed negative relationships with six of the remaining eight statements in the path. No fewer than 42 of the 110 statements correlated negatively with at least one other statement from the same path.

By eliminating these 42 statements (as many as seven from a single path) each of the 13 matrices was reduced to a positive manifold. Particularly vulnerable to this type of instrument sharpening were statements negating or rejecting something, as, for example, *Excessive desires should be avoided and moderation sought,* and *A person should not hold on to himself, withdraw from people, keep aloof and self-centered.* While of the total of 110 statements, 35 per cent were negative assertions, in the eliminated group of 42, 62 per cent were negations, a disproportion that is highly significant, χ^2 (1) $= 20.32, p < .01$.

In order to assess the effect of the eliminations on the homogeneity of the revised paragraphs, Phase 2 was completed: 40 *Ss* rated and ranked the revised and original forms and made Q sorts of the 110 items. Within-path intercorrelation matrices were derived for the revised form in the manner previously described

for the original. The mean within-path r's of the statements in the original and revised forms were then computed, using the matrices in Phase 1 for the former and those in Phase 2 for the latter. These results are contained in Columns A and B of Table 1.

Clearly the homogeneity of those paths from which discordant elements were eliminated (Paths 1-11) is improved, the mean r for these in the original form being .13, that for the revised, .22. On an r to z transformation the mean differences in the two sets of r's is highly significant, $d_s = .09$, $t\ (10) = 3.46$, $p < .01$. Despite this marked improvement in homogeneity, the within-path relationships of some of the paragraphs are still very low, e.g., for Path 7, the mean inter-item r is only .09. It should be noted that homogeneity, in the sense used above, entails no necessary restriction on the diversity of *stimulus items* comprising a given path. It is rather an expression of the unity or integration with which subjects respond to the path's varied content.

TABLE 1

Mean Within-Path Part-Part[a] Correlations (r's) for Three Versions of Morris's Ways to Live

Path	Normative Sample ($N = 24$)	Cross-validating Sample ($N = 40$)	
	A Original Form	B Revised Form	C Short Form
Path 1	.09	.11	.19
Path 2	.17	.30	.36
Path 3	.15	.27	.37
Path 4	.04	.27	.67
Path 5	.20	.18	.43
Path 6	.21	.26	.38
Path 7	.04	.09	.06
Path 8	.14	.16	...[c]
Path 9	.02	.12	−.01
Path 10	.06	.27	.40
Path 11	.31	.34	.31
Path 12	.39[b]	.29[b]	.41
Path 13	.38[b]	.25[b]	.61
Mean of path means:			
Paths 1–11	.13	.22	...[c]
All but Path 8	.18	.24	.36
Paths 1–13	.17	.23	...[c]

[a] Q sort values of statements comprising paths.
[b] Original and revised forms identical for these paths.
[c] In the short form Path 8 a single statement.

2. The Question of Simplicity.

In the process of constructing the revised form the homogeneity of the paths was increased without deliberate attempt to reduce their complexity as stimuli. Ranking them and rating them still require a high level of synthesis. The second question raised in this study concerns the simplification of the paths, the reduction of the number of details to be processed in the task of evaluating them. The proposed solution rests on the assumption that each of the original paths, even the most heterogeneous, contains some shorter expression that may be taken as a statement of its basic significance or essence. If the existence of such key expressions can be demonstrated empirically, it should be possible to produce an efficient short form of the "Ways to Live" document.

Two methods suggest themselves for isolating key expressions in the paths. The first is simply to select as representative of a given path the statement that shows the highest average correlation with all the statements in the original paragraph. The second method is to select for each path the statement to which *Ss* respond most nearly in the way they respond to the paragraph as a whole. The first of these approaches entails the analysis of part-part relationships, and the second of part-whole. There is no reason to expect the two methods to yield identical results, nor indeed to view either as intrinsically superior. In the present study both methods are used.

The 13 matrices calculated in Phase 1 were used in the selection of key expressions. Within each path the statements were ranked in terms of the size of their average intercorrelations. The highest ranking statement in each path was automatically included in the short form, plus as many additional statements as were needed to make the brief path descriptions sufficiently detailed to be distinctive and roughly equivalent in length. Path 8 is composed of one long statement, Path 2 of four short ones; the remaining paths contain either two or three. Morris's original wording was adhered to as closely as possible. The short form, brief enough to be reproduced on a single 8½ × 11 sheet, is shown in Figure 1.

As in the revised form, the items in the short form were cross-validated on the 40 *Ss* in Phase 2. To check part-part relationships in this form, the appropriate inter-item coefficients were se-

FIGURE 1

"WAYS TO LIVE"

On the following page are described 13 ways to live, which various persons at various times have advocated and followed. In the left margin rank these ways *in the order you prefer them,* so that the number *1* is by the path you like best, the number *2* by that you like next best, and so on, with number *13* by the path you like least.

It is not a question of what kind of life you now lead, or the kind of life you think it prudent to live in our society, or the kind of life you think good for other persons, *but simply the kind of life you personally would like to live.*

——— *Path 1.* An individual should actively participate in the social life of his community, not to change it primarily, but to understand, appreciate, and preserve the best that man has attained. Life should have clarity, balance, refinement, control.

——— *Path 2.* The individual should for the most part "go it alone," having much time to himself, stressing self-sufficiency, reflection and meditation, knowledge of himself. The center of life should be found within the self.

——— *Path 3.* This way of life makes central the sympathetic concern for other persons. Whatever hinders sympathetic love among persons should be avoided, for such love alone gives significance to life. One should become receptive, appreciative, and helpful with respect to others.

——— *Path 4.* Life should be more a festival than a workshop, or a school for moral discipline; it should be enjoyed, sensuously enjoyed, enjoyed with relish and abandonment. To let oneself go, to let things and persons affect oneself, is more important than to do—or to do good.

——— *Path 5.* A person should merge himself with a social group, enjoy cooperation and companionship, join with others in resolute activity for the realization of common goals. Life should merge energetic group activity and cooperative group enjoyment.

——— *Path 6.* We should stress the realistic solution of specific problems as they appear and the improvement of techniques for controlling the world and society. We have to work resolutely and continually if control is to be gained over the forces that threaten us.

——— *Path 7.* We should at various times and in various ways accept something from all other paths of life, but give no one our exclusive allegiance. We must cultivate flexibility, admit diversity in ourselves, accepting the tension which this diversity produces.

——— *Path 8.* The enjoyment of simple, easily obtainable pleasures should be the keynote of life: the pleasures of just existing, of savory food, of comfortable surroundings, of talking with friends, of rest, relaxation.

——— *Path 9.* The good things in life cannot be found by resolute action, or by participation in the turmoil of social life. One should cease to make demands, waiting in quiet receptivity, open to the powers which nourish the self and work through it. Sustained by these powers, one knows joy and peace.

——— *Path 10.* Self-control should be the keynote of life, not the easy self-control which retreats from the world, but the vigilant, stern, manly control of a self that lives in the world. One should hold firm to high ideals, and not be bent by the seductive voices of comfort and desire.

——— *Path 11.* The contemplative life is the good life, the life that is rewarding. The rich internal world of ideals, of sensitive feelings, of reverie, of self-knowledge is man's true home.

Path 12. The use of the body's energy is the secret of a rewarding life. Not in cautious foresight, not in relaxed ease does life attain fulfillment,

for it is the active deed that is satisfying, the deed adequate to the present, the daring and adventurous deed.

Path 13. One should let himself be used by other persons in their growth, and by the great objective purposes in the universe. One should be a serene, confident, quiet vessel, guided by the great dependable powers which silently and irresistibly achieve their goal.

lected from the within-path matrices, and the mean r determined for each path (Table 1, Column C). The mean within-path correlation of the short form is .36, a figure that compares favorably with a corresponding mean of .18 for the original form and .24 for the revised form (Table 1). Just as the homogeneity of the revised form was significantly greater than that of the original, the homogeneity of the short form exceeds that of the revised, $d_s = .14$, t $(11) = 2.50$, $p < .05$. In terms of part-part relationships, then, the paragraphs of the short form appear to represent a more consistent core of meaning than either the original or the revised form. Despite this overall improvement, however, it is to be noted that two paths (7 and 9) are still as heterogeneous as they were originally.

To examine part-whole relationships, the Q sort values for each of the 110 items were correlated with the 40 *Ss'* ratings of the original paths. The range of these 110 r's was large, from $-.33$ to $+.67$. As in the analysis of part-part relationships, three separate mean r's were calculated for each path, one based on all the items in the original paragraph, a second on the items in the revised path, and a third on the items in the short path. These means are presented in Columns A, B, and C, respectively of Table 2. Each progressive change in the composition of the paths is accompanied by a highly significant improvement in mean part-whole relationships. For the 11 paths modified from the original to the revised form, the mean r rises from .23 to .30, $d_s = .08$, t $(10) = 6.67$, $p < .01$. All 13 paths were altered from the revised to the short form and the comparable figures are .31 for the revised and .39 for the short, $d_s = .09$, t $(12) = 3.91$, $p < .01$. Again it should be noted that even with the progressive improvement in part-whole r's, Paths 7 and 9 remain, as in the part-part analysis among the least homogenous paths.

The reliable improvement in both part-part and part-whole relationships indicates that the paragraphs of the short form taken

individually express the core conceptions of Morris's document more adequately than the original paragraphs themselves.

TABLE 2

Mean Within-Path Part-Whole[a] Correlations (r's) for Three Versions of Morris's Ways to Live as Shown in Cross-Validating Sample (N = 40)

Path	A Original Form	B Revised Form	C Short Form
Path 1	.26	.32	.46
Path 2	.29	.39	.50
Path 3	.24	.34	.42
Path 4	.17	.31	.38
Path 5	.22	.34	.45
Path 6	.17	.23	.14
Path 7	.12	.19	.27
Path 8	.34	.36	.48
Path 9	.10	.17	.18
Path 10	.26	.32	.35
Path 11	.26	.30	.35
Path 12	.44[b]	.44[b]	.48
Path 13	.33[b]	.33[b]	.52
Mean of path means:			
Paths 1–11	.23	.30	.36
Paths 1–13	.25	.31	.39

[a] Part = (Q sort values of statements comprising paths) vs. Whole = (Ratings of original paragraphs).
[b] Original and revised forms identical for these paths.

3. *The Question of Reliability.*

While the complexes of items comprising the revised and short forms have been shown to be more homogeneous and simpler than those of the original, the reliability of each form as a separate document has not yet been explored. In the final phase of the present study estimates of test-retest and alternate-form relationships were obtained both across paths and across individuals, special attention being paid to the short form.

Some of the reliability coefficients reported by Morris (1956) appear to be across-path estimates, a form of estimate suitable for the group comparisons he was concerned with, but one that is hardly appropriate for the study of individual differences. Using *Ss'* ratings, Morris estimates the repeat reliability of his document to be about .85 for college students; he reports average coefficients of .87 and .78 from two specific studies. The across-path reliability coefficients reported in the present study are Spearman rho's

based on mean ranks of the paths as shown in six independent samples ranging in size from 32 to 46 *S*s. For each form two samples were used, thus making available a single repeat reliability estimate for each form and four alternate-form estimates for each pair of them. Repeat estimates are uniformly high, .93, .95, and .96 for the original, revised, and short forms, respectively. The mean alternate form estimates are lower: .83 between original and revised forms, .82 between revised and short, and .71 between original and short. These lower coefficients reflect changes in the overall acceptability primarily of two paths as a consequence of the refining process. The contemplative life (Path 11) becomes more attractive with the elimination of statements like *The external world is no fit habitat for man; it is too big, too cold, too pressing.* Conversely, the enjoyment of simple pleasures (Path 8) becomes relatively less attractive when the discordant elements have been removed from the more active, social, and achievement-oriented paths.

The analysis of across-individual relationships also reveals a relatively high level of consistency. For each person who took two forms of the document, or the short form twice, a Spearman rho was calculated from his two rankings. On the basis of z's, mean rho's were as follows: for the short-short sample the mean rho was .80 ($N = 32$); for the short-revised sample, .72 ($N = 35$); for the revised-original sample, .63 ($N = 40$); and for the short-original sample, .57 ($N = 35$). The mean rho of .80 is somewhat higher than the reliability estimate of .67 (mean Pearsonian r based on individual repeat *ratings*) reported by Morris and Jones (1955).

On the basis of the mean test-retest coefficient of .80 it is evident that the path descriptions in the short form are sufficiently distinctive to be ranked reliably. This, of course, does not imply that every *S*'s performance is consistent; in each of these samples a few individual coefficients were at the zero level. Since inconsistent performers cannot be identified on the basis of a single administration, it is recommended that repeat rankings be obtained whenever individual performances are a matter of concern.

The mean alternate-form coefficient of .72 between the short and revised forms was not significantly different from the repeat reliability of the short form itself, $d_z = 20$, $t\ (65) = 1.67$, $p > .05$.

Within the limits of these reliabilities, these forms may be taken as equivalent. With the original form, however, the short form shows a mean coefficient (.57) that is significantly lower than its repeat reliability, $d_s = .46$, $t\ (65) = 3.83$, $p < .01$. Thus, while there is a substantial relationship between the original and short forms, there are also important differences between them. As with the across-path results, these differences reflect systematic changes in the acceptability of a few paths following the elimination of certain statements. The differences noted earlier for Path 11, the contemplative path, also appear on the individual level, for example, and the place of enjoyment is again a focus for down-grading. This time, however, the most obvious shifts are in the reduced acceptability of festivity and sensuous enjoyment (Path 4) rather than of the simple, easily obtainable pleasures (Path 8).

A comparison of the content of the short form with that of the original indicates that Morris's *a priori* judgment as to what was central in the various paths was remarkably accurate. His introductory statements or topic sentences in ten of the 13 original descriptions emerged as key expressions from the empirical analysis as well. Between Morris's brief characterizations of the paths and the implications of the data of this study, however, some notable differences do occur. Path 4, for example, Morris believed to emphasize festivity and solitude in alternation. Analytically it emphasizes festivity and sensuous enjoyment, solitude being irrelevant. In the present analysis Path 7 appears to be purely eclectic, emphasizing diversity and reflecting not at all the integration Morris posited as central. Also, in Path 12, the use of the body's energy plays a far more important part empirically than Morris's characterization of it as chancing adventuresome deeds would imply. These and other less notable differences suggest the following as a more accurate epitome of the paths than that given by Morris (1956) in his book and cited in the opening paragraph of this paper (alterations in italics):

Way 1: *Appreciate and* preserve the best man has attained.
Way 2: Cultivate independence *and self-knowledge.*
Way 3: Show sympathetic concern for others.
Way 4: Experience festivity *and sensuous enjoyment.*
Way 5: Act and enjoy life through group participation.

Way 6: Master *threatening* forces by constant *practical work.*
Way 7: *Admit diversity and accept something from all ways of life.*
Way 8: *Enjoy the simple, easily obtainable pleasures.*
Way 9: Wait in quiet receptivity *for joy and peace.*
Way 10: Control the self *and hold firm to high ideals.*
Way 11: Meditate on the inner life.
Way 12: *Use the body's energy* in daring and adventurous deeds.
Way 13: *Let oneself be used by* the great cosmic purposes.

Summary

The 13 paragraphs of Morris's document were divided into 110 statements, Q sorts of which were used to assess the coherence of his complex stimulus-situations. Eleven paths showed discordancy (negative within-path inter-item *r*'s), the bulk of the negative relationships stemming from "thou shalt not" statements rather than from assertations of positive value. A *revised* form (discordant statements omitted) showed improved homogeneity in cross-validation, and a *short* form (composed of empirically-derived key expressions) showed still higher homogeneity. Across-path reliability is high in all forms. Alternate-form comparisons indicate equivalence of short and revised forms, with some systematic changes from the original. Test-retest reliability of the short form is .80. *S*s were 230 college students.

REFERENCES

Morris, C. *Varieties of Human Value.* Chicago: University of Chicago Press, 1956.

Morris, C. and Jones, L. V. Value Scales and Dimensions. *Journal of Abnormal and Social Psychology*, 1955, 51, 523-535.

Osgood, C. E., Ware, E. E., and Morris, C. Analysis of the Connotative Meanings of a Variety of Human Values as Expressed by American College Students. *Journal of Abnormal and Social Psychology*, 1961, 62, 62-73.

Winthrop, H. Psychology and Value: A Critique of Morris's Approach to Evaluation as Behavior. *Journal of General Psychology*, 1959, 61, 13-37.

Educational and Psychological Measurement
1966, 26, 883–886.

TEST-RETEST RELIABILITY OF THE EPPS[1]

DANIEL V. CAPUTO
Queens College (CUNY)
GEORGE PSATHAS and JON M. PLAPP
Washington University, St. Louis

In his manual for the use of the Edwards Personal Preference Schedule, (EPPS), Edwards (1959) reported test-retest relibility coefficients for each of the 15 need scores plus a consistency score for a group of 89 University of Washington students who were retested after a one week interval. Subsequent studies have reported test-retest reliabilities of need scores after intervals of three weeks (Mann, 1958) and one week (Horst and Wright, 1959). Coefficients from these studies are reported in Table 1. Borislow (1958) has also reported test-retest reliabilities ranging from .65 to .91 for the *profile* of need scores for *S*s after a two week interval.

All these studies have demonstrated satisfactory short-term reliability for the test and consequently call into question the argument of Levonian, Comrey, Levy, and Procter (1959) who stated that the basic form of the EPPS item is one that encourages low reliability of response since the *S* must choose which of two statements seems more descriptive of himself while at the same time this choice is made more difficult by equating the statements for social desirability. This procedure maximizes the number of difficult and hence potentially unreliable choices for the *S*.

As part of a broader study of the EPPS used with nursing students, it is possible to report reliability coefficients for a much longer inter-test period than previously investigated. The EPPS was administered to a group of 79 female freshman nursing stu-

[1] This paper is a partial report of a research project "Role Differentials and Nursing Ideology," John A. Stern and Albert F. Wessen, Co-Principal Investigators, PHS Grant NU-00050. A portion of the computation was done with support provided by the Washington University Computer Center under National Science Foundation Grant G-22296.

TABLE 1

Test-Retest Correlations at Different Intervals

Need	Edwards (1 wk. interval)			Horst & Wright (1 wk. interval)	Mann (3 wk. interval)	Present Study (15 month interval)				
	r_{12}	$\bar{X}_1$	SD_1	r_{12}	r_{12}	r_{12}	$\bar{X}_1$	$\bar{X}_2$	SD_1	SD_2
Ach	.74	14.46	4.09	.83	.64	.47[a]	11.38	11.85	4.02	3.94
Def	.78	12.02	3.68	.67	.87	.45[a]	12.10[d]	10.85[d]	3.00	3.43
Ord	.87	11.31	4.45	.81	.77	.59[a]	9.42	9.71	4.32	4.14
Exh	.74	14.43	3.67	.64	.71	.60	15.17	15.44	3.68	3.85
Aut	.83	13.62	4.48	.76	.76	.65[b]	11.42	12.06	4.48	4.27
Aff	.77	15.40	4.09	.80	.55	.64	16.73[e]	15.71[e]	3.81	4.26
Int	.86	17.00	5.60	.84	.67	.41[a]	19.11	19.10	4.51	4.78
Suc	.78	12.09	4.59	.78	.72	.46[a]	13.75	12.77	4.31	4.21
Dom	.87	15.72	5.28	.82	.73	.60[a]	11.35[e]	12.96[e]	4.64	4.60
Aba	.88	14.10	4.96	.84	.69	.59[a]	17.75	16.69	4.69	4.77
Nur	.79	14.04	4.78	.81	.59	.65	18.03[e]	15.56[e]	4.95	4.19
Chg	.83	16.17	4.88	.80	.86	.73	17.48	17.67	5.02	4.73
End	.86	12.52	5.11	.76	.77	.61[a]	11.92	11.00	4.82	4.23
Het	.85	15.08	5.66	.76	.85	.45[a]	14.31[e]	16.83[e]	5.02	4.85
Agg	.78	11.55	4.57	.72	.80	.50[a]	10.13	11.21	3.65	4.67
Consistency	.78	11.59	1.78	—	—	—	—	—	—	—
N	89			92	96	52				

[a] Present correlation significantly smaller than Edwards corresponding correlation at .01 level.
[b] Present correlation significantly smaller than Edwards corresponding correlation at .02 level.
[c] Means significantly different at .01 level.
[d] Means significantly different at .02 level.
[e] Means significantly different at .05 level.

dents in November, 1962, and readministered to the 52 students who were still in school in March, 1964. Test-retest product-moment correlations were computed for each of the EPPS needs across the 52 *Ss* who had taken the test at both administrations. These coefficients and the corresponding means and standard deviations are reported in Table 1. Profile intercorrelations were also determined for each student by correlating her profile of 15 need scores at time 1 with her profile at time 2. The distribution of these correlation coefficients is shown in Figure 1.

The test-retest correlations between needs in the present study are all significantly greater than zero. However, all are lower than those obtained in the studies which used a one-week interval (see Table 1). Use of *z*-tests to compare the correlations of the present study with those of Edwards' reliability study showed that 11 of 15 of the present correlations are significantly lower. With the exception of Affiliation and Nurturance, whose test-retest correlations are higher in the present study, all are also lower than those obtained by Mann in his three-week interval study. Thus, the correlations between *Ss*' scores on the individual EPPS needs before and after a 15-month interval are significantly positive though not as high as are comparable correlations obtained using shorter test-retest intervals.

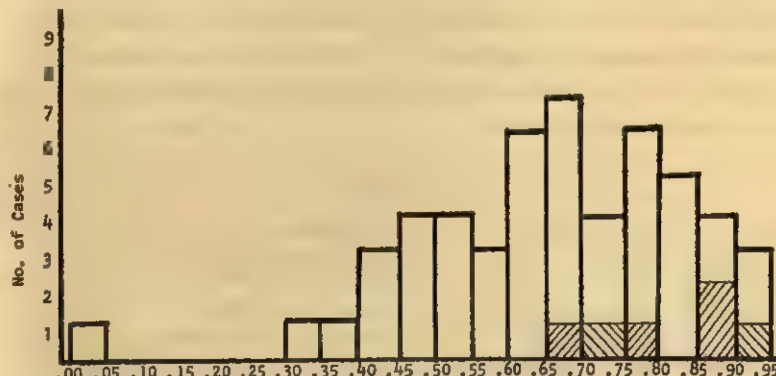


Figure 1. Distribution of individual profile correlations all 15 needs, Time 1 with Time 2.

Borislow's six reported correlations

$r = .273, p = .05$

$r = .354, p = .01$

Examination of the correlations obtained between the individual profiles of *S*s before and after the 15-month interval (see Figure 1) reveals an approximately normal distribution. The mean of the 52 profile correlations is .66, and is significantly greater than zero ($p < .05$). Fifty-one of the profile correlations are significantly greater than zero at the .05 level, while 48 of these also attain significance at the .01 level. The one *S* whose profile correlation ($r = .03$) is not significant obtained a Consistency score of 9/15 on the second test administration which suggests that her responses at that time may have been made randomly (Edwards, 1959). Borislow's individual profile correlations, obtained using a two-week test-retest interval, are also shown in Figure 1. Five of the six are greater than the mean of the present profile correlations.

The possibility of significant differences existing between the pre and post 15-month interval need scores of the subjects of the present study was investigated by applying *t*-tests for correlated means to the need scores at time 1 and time 2. As shown in Table 1, needs Deference, Affiliation, and Nurturance were significantly lower at the second testing time, while needs Dominance and Heterosexuality were significantly higher.

The question may be raised as to whether a retest after a long time interval such as the present one represents an investigation of test reliability or subject change. Obviously, both may be involved. However, from the perspective of test reliability it can be concluded that over the long term, the EPPS shows acceptable reliability both in regard to measurement of a single scale over all *S*s and for a single *S* over all scales.

REFERENCES

Borislow, B. The Edwards Personal Preference Schedule (EPPS) and Fakability. *Journal of Applied Psychology*, 1958, 42, 22-27.

Edwards, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corp., 1959.

Horst, P. and Wright, C. E. The Comparative Reliability of Two Techniques of Personality Appraisal. *Journal of Clinical Psychology*, 1959, 15, 388-391.

Levonian, E., Comrey, A., Levy, W., and Procter, D. A Statistical Evaluation of Edwards Personal Preference Schedule. *Journal of Applied Psychology*, 1959, 43, 355-359.

Mann, J. H. Self-ratings and the EPPS. *Journal of Applied Psychology*, 1958, 42, 267-268.

Educational and Psychological Measurement
1966, 26, 887–893.

CERTAIN CONSEQUENCES OF APPLYING THE *K* FACTOR TO MMPI SCORES

GEORGE D. YONGE
University of California, Davis

The present study is focused on some of the psychometric consequences of using the *K* scale to correct certain *Minnesota Multiphasic Personality Inventory* (MMPI) clinical scales for distortions induced by test taking attitudes.

There is a growing bibliography documenting the ineffectiveness of using the *K* correction as a means for improving the clinical validity of certain MMPI scales (see Dicken, 1963). At the same time, it is instructive to note that the psychometric consequences of using the *K* scale as a correction factor have seldom been evaluated apart from focusing on external criteria. This is not surprising in that the philosophy underlying empirically derived tests, such as the MMPI, places a premium on the ability of the test to correlate with a criterion. Any question of a logical or psychological nature is clearly subordinate to this concern for correlations with criteria. A clear statement of this philosophy is provided by Fricke (1963) who essentially states that the characteristics of a test—reliability, content, etc.—are not important as long as the test predicts the criterion. There would be no argument with this philosophy if the test has a high correlation with a criterion (which is seldom the case) *and* if the psychologist is content to make no psychological interpretations of the test scores.

It is contended that most psychologists use tests such as the MMPI to learn something about individuals rather than to blindly predict criterion scores or group membership. To the extent that this contention is true, questions must be asked regarding the logical and psychological aspects of the measurement operations performed.

When the use of the K score as a correction factor is evaluated from a logical, psychological, and psychometric perspective, there is little justification for its use.

One study which raises some psychometric questions in connection with the use of K is that by Tyler and Michaelis (1953). These authors formulated six questions to be considered in evaluating the use of the K scale as a correction for response distortion. Whereas Tyler and Michaelis addressed themselves to four of these questions, the present study is addressed to two of the four treated by Tyler and Michaelis and one not asked by them.

Specifically, the questions[1] asked in the present study are:

1. What are the correlations between the K scores and scores on the clinical scales to which K is applicable?
2. What are the correlations between the K-corrected and uncorrected scores?
3. Does the K correction affect the correlations of the MMPI scales with other measures of social-emotional adjustment?

Method

One hundred and thirty-six[2] University of California students applying for a study abroad program were given the *Omnibus Personality Inventory* (Center for the Study of Higher Education, 1962) and the *Minnesota Multiphasic Personality Inventory* (Hathaway and McKinley, 1951). The answers to the above three questions were approached by means of Pearson product-moment correlations. Although not reported, the essential results of the present study were also obtained with a sample of 58 University of California Peace Corps applicants.

Results

1. *What are the correlations between K scores and the scores on the clinical scales to which K is applicable?*

The answer to this question, in part, underlies the answers to the subsequent questions. Table 1 presents the correlations between the K scores and the scores derived from the relevant clinical scales.

[1] Questions 1 and 2 were raised and treated by Tyler and Michaelis; question 3 is unique to the present study.

[2] The data on which the present study is based were obtained from Mrs. Barbara Kirk, University of California Counseling Center.

For comparative purposes, the results obtained by Tyler and Michaelis are also reported.

The correlations between *K* and the relevant clinical scores are remarkably similar to those obtained by Tyler and Michaelis. It is also interesting to note that there is a rough correspondence between the magnitude of these correlations and the proportion of *K* recommended to correct a given scale.[3] It is curious, however, that no correction is recommended for the Hysteria (*Hy*) scale, which correlates .39 with *K*, or for the Social Introversion (*Si*) scale, which correlates —.55 with *K*. If indeed *K* is a measure of response bias, *Hy* and *Si* are about as permeated with this unwanted variance as are the scales for which a *K* correlation is recommended.

2. *What are the correlations between the K-corrected and uncorrected scores?*

Table 2 presents the data relevant to this question. Again, for comparative purposes, the results obtained by Tyler and Michaelis are also reported.

The high correlations between the corrected and uncorrected

TABLE 1

Correlations between K and the MMPI Scales to Which K is Applicable

	MMPI Scales					
	N	*Hs*	*Pd*	*Pt*	*Sc*	*Ma*
Present Study	136	−.41*	−.28*	−.75*	−.63*	−.39*
Tyler and Michaelis	56	−.46*	−.24	−.69*	−.59*	−.37*

* Significant .01 level.

TABLE 2

Correlations between K-Corrected and Uncorrected MMPI Scores

Correlated Scores	Present Study ($N = 136$)	Tyler and Michaelis ($N = 56$)
Hs · *Hs* + .5*K*	.55*	.66*
Pd · *Pd* + .4*K*	.90*	.91*
Pt · *Pt* + *K*	.64*	.67*
Sc · *Sc* + *K*	.58*	.56*
Ma · *Ma* + .2*K*	.98*	.97*

* Significant .01 level.

[3] The recommended proportions of *K* are: *Hs* + .5*K*, *Pd* + .4*K*, *Pt* + 1.0*K*, *Sc* + 7.0*K*, and *Ma* + .2*K*.

scores for Psychopathic Deviate (*Pd*) and Hypomania (*Ma*) indicate that the *K* correction has a minor effect upon these two scales. This finding is not surprising and is no doubt a reflection of the low correlations these scales have with *K* and the small proportion of *K* used to correct these scales. Tyler and Michaelis, noting that the reported reliabilities of *Pd* and *Ma* are considerably lower than the corresponding corrected-uncorrected correlations, concluded that the uncorrected scores are apparently better measures of the corrected scores than of themselves. Unfortunately, Tyler and Michaelis have overlooked the spurious nature of these correlations owing to the part-whole correlation involved. The spurious nature of the correlations presented in Table 2 is a strong recommendation against comparing them with the reliabilities of the scales.

The point to be made from the correlations in Table 2 is that the *K* correction markedly alters whatever is being measured by Hypochondriasis (*Ha*), Psychasthenia (*Pt*), and Schizophrenia (*Sc*). Furthermore, these correlations underplay the change of measurement which has taken place simply because they are spurious correlations.

A consideration of the data presented in Table 2 gives rise to a question asked but not treated by Tyler and Michaelis; namely, how do the reliabilities of the corrected scores compare with those of the uncorrected scores? Since this question can be treated only in a tangential manner in the present study, it was not raised as a major question.

Kuder-Richardson Formula 21 reliability coefficients were computed for the corrected and uncorrected *Pt* and *Sc* scores. These two scales were used for this analysis because the entire *K* score is added to each of these scores as the recommended correction. Thus, by correction, the length of these scales is increased by approximately 30 items.[4] However, as indicated in Table 3, the Kuder-Richardson reliabilities change from high positive to low negative values. Thus, from the point of view of internal consistency reliability, the measurement attained with the uncorrected scales is negated or cancelled out when the *K* correction is applied. It should be added that the data in Table 3 offer no implications regarding test-retest reliabilities. In fact, Rosen (1953) has reported that

[4] The item overlap between *K* and *Pt* is two, between *K* and *Sc* one.

with a sample of 40 psychiatric patients retested after four days (on the average) the test-retest reliabilities of *Pt* and *Sc* increased slightly when the *K* correction was used.

3. *Does the K correction affect the correlation with other measures of social-emotional adjustment?*

It will be remembered that the subjects in the present study were also administered the *Omnibus Personality Inventory* (OPI). There are three measures in the OPI which presumably focus on social-emotional adjustment. These are Impulse Expression (*IE*), Schizoid Functioning (*SF*), and Lack of Anxiety (*LA*). The correlations between these OPI measures and the corrected and uncorrected MMPI scores are presented in Table 4.

The most striking feature of the data presented in Table 4 is the consistent reduction in the OPI-MMPI correlations when the *K*

TABLE 3

Means, Standard Deviations, and Kuder-Richardson Formula 21 Reliability Coefficients of K-Corrected and Uncorrected Psychasthenia and Schizophrenia Scores

	Uncorrected			Corrected		
	KR 21	Mean	S.D.	KR 21	Mean	S.D.
Pt	.78	8.5	5.5	−.28	26.5	3.7
Sc	.73	8.1	5.1	−.18	26.2	4.1

TABLE 4

Correlations between Selected OPI Social-Emotional Adjustment Scales and K-Corrected and Uncorrected MMPI Scales

	OPI Scales		
MMPI Scales	*IE*	*SF*	*LA*
Hs	.14	.40*	−.37*
Hs + .5*K*	−.26*	−.31*	.20*
Pd	.51*	.42*	−.42*
Pd + .4*K*	.33*	.09	−.16
Pt	.44*	.82*	−.80*
Pt + *K*	.16	.37*	−.52*
Sc	.60*	.74*	−.66*
Sc + *K*	.31*	.17	−.21*
Ma	.47*	.41*	−.30*
Ma + .2*K*	.41*	.27*	−.19
K	−.43*	−.74*	.59*

* Significant .01 level.

correction is applied. Particularly, the reduction in the *Pt-SF*, *Pt-LA*, *Sc-SF*, and *Sc-LA* correlations may be interpreted as indicating that the *K* factor reduces the validity of these MMPI scales. This consequence of using the *K* correction is, of course, consistent with the other data presented in the present paper.

Discussion and Conclusions

As a measure of test-taking attitude, the *K* scale represents a concerted effort to assess a person's defensiveness or tendency to place himself in a good light. In view of the socially undesirable content of the clinical scales, the defensiveness reflected in the *K* scale will presumably "hold down" a person's scores on the clinical scales. Therefore, by adding *K* (Defensiveness) to *Pt*, for example, the "defensiveness" variance in *Pt* will be cancelled out (since *K* and *Pt* correlate negatively, adding *K* amounts to a subtraction or reduction of variance). This assumes that Psychasthenia and defensiveness are additive components present in the uncorrected score. A parallel argument in the area of achievement testing would have the psychometrist add to the raw score a measure of a tendency to make careless errors in order to obtain a more accurate estimate of the subject's level of achievement. Since it is recognized that errors, careless or otherwise, are aspects of the person's achievement status, we do not find this operation applicable to the achievement domain. And, in fact, if a measure of poor achievement (errors) were added to a measure of good achievement (correct responses), it would seem that the content of the measure would be washed out rather than purified. The data presented in the present study are interpreted as indicating that the *K* correction nullifies rather than purifies the measurement obtained with the uncorrected scales.

Within the perspective of the above discussion, a further consideration of the *Pt* and *Sc* scales, to which the entire *K* score is added as a correction, will point to the unusual logic and consequences connected with the use of the *K* factor.

Adding *K* to *Pt* and to *Sc*, where the correlations are —.75 and —.63, respectively, is tantamount to adding opposites and reasoning that this procedure will purify the *Pt* and *Sc* dimensions. Operationally, this is similar to adding correct and incorrect answers

on an achievement test; this would tend to reduce the scale variance and virtually cancel out the content of the test. By adding *K* to *Pt* and to *Sc* (i.e., increasing their lengths by 30 items) their respective standard deviations are *decreased* from 5.5 to 3.7 and from 5.1 to 4.1! This is analogous to what would be expected if correct and incorrect answers to an achievement test were added together.

When there is evidence that a person's scores are influenced by test taking attitudes, it is difficult to see how intra-test manipulations of these biased scores will somehow rid them of their bias. Rather, it would seem more reasonable to assess the person's personality, attitudes, etc., by some *other* measurement operation.

The array of data considered in the present paper converge on one general conclusion. The application of the *K* factor, from a psychometric viewpoint, considerably alters the structure of the MMPI scales. Specifically, the *K* correction reduces the internal consistency reliability and the validity of these measures. The above findings, in conjunction with the growing body of negative evidence concerning the utility of the *K* correction in increasing the clinical validity of the MMPI, clearly argue against the continued use of *K* as a correction factor.

REFERENCES

Center for the Study of Higher Education. *Omnibus Personality Inventory:* Research Manual. Berkeley: Center for the Study of Higher Education, 1962.

Dicken, C. Good Impression, Social Desirability, and Acquiescence as Suppressor Variables. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 699-720.

Fricke, B. G. *Opinion, Attitude, and Interest Survey Handbook: A Guide to Personality and Interest Measurement.* Ann Arbor: University of Michigan, Evaluation and Examination Division, 1963.

Hathaway, S. R. and McKinley, J. C. *Minnesota Multiphasic Personality Inventory: Manual.* New York: Psychological Corporation, 1951.

Rosen, A. Test-retest Stability of the MMPI. *Journal of Clinical Psychology,* 1953, 17, 217-221.

Tyler, F. T. and Michaelis, J. U. *K*-Scores Applied to MMPI Scales for College Women. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1953, 12, 459-466.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 895–1072.

VALIDITY STUDIES SECTION

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

IMPORTANT NOTICE TO AUTHORS

THE VALIDITY STUDIES SECTION is published twice a year, once in the Summer issue and again in the Winter issue, for which the closing dates for receiving manuscripts are February first and August first, respectively. Although articles between two and eight printed pages are usually preferred, an occasional exception is made to publish articles of somewhat greater length.

Considerable flexibility exists concerning format as can be seen from a study of recently published articles. However, the model presented in the Spring, 1953, issue of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT still represents a close approximation to what is customarily published. The prospective contributor is encouraged to read the original announcement.

In order that the usual number of articles of other types may not be reduced, it is necessary to enlarge the journal and to charge the authors for most of the publishing costs. Due to the financial pressure of rising printing costs, the new rate for publication of articles in this section, beginning with the first issue for the year of 1967, will be twenty-five dollars per page. The extra cost of the composition of tables and formulas will be added to the basic rate. One hundred gratis reprints is extended to the author of articles appearing in this journal.

Two copies of manuscripts should be sent to:

Dr. William B. Michael
Professor of Education and Psychology
University of California, Santa Barbara
Santa Barbara, California

Educational and Psychological Measurement
1966, 26, 899–915.

NONINTELLECTIVE PREDICTORS OF ACHIEVEMENT IN COLLEGE[1]

ROBERT C. NICHOLS
National Merit Scholarship Corporation

Selection of students in college admissions and scholarship programs is usually based in part on a prediction of the student's performance in college. The accepted predictors are aptitude test scores and high school grades, often combined into a single index by a regression equation. The conventional criterion for validation of these predictions is the college grade point average, usually for the freshman year. There is a well developed technology for making these predictions, and many colleges and scholarship agencies have developed formulas for optimum weighting of high school grades and test scores in their local situation.

The growing uniformity of selection practices might lead one to assume that the prediction of college performance has reached a state of stable maturity satisfactory to most practitioners. Instead, there is increasing concern and dissatisfaction with the current state of the art. This concern may be, in part, a reaction to the recently vocal lay critics of the testing movement, but there is also a recognition of the inadequacy of current selection methods to meet the demands being made of them. There are two major sources of dissatisfaction:

1. The popular colleges and most scholarship and honors programs have many more applicants than they can accept. After applicants with low grades or low test scores are eliminated, further discriminations must be made between the remaining highly quali-

[1] This study is part of the research program of the National Merit Scholarship Corporation and was supported by grants from the Ford Foundation, the Carnegie Corporation of New York, and the National Science Foundation.

fied candidates. The selection committee can make these decisions on the basis of the remaining small differences in test scores and high school rank; they can look at other data, such as an interviewer's impression, a recommendation or an autobiography; or they can flip a coin. These three methods of discriminating between students with good grades and high test scores are about equally valid when judged by the criterion of later student performance. Educational researchers have been busy searching for indicators of success other than grades and test scores (usually called non-intellective predictors) and some have been found. However, none has yet achieved enough demonstrated success to be widely adopted.

2. There is growing dissatisfaction with the use of college grades as the criterion for evaluation of predictors. Some educators are beginning to feel that the student who makes the best grades is not necessarily the most valuable student. For example, Stalnaker (1965) in discussing the National Merit Scholarship selection program, said, ". . . we want to find students who will succeed in college, but—much more important—will also use their college education in some socially desirable, productive way after graduation. How relevant are grades to this goal? . . . Do you inquire of your accountant, your physician or your lawyer the grades he received in college? Predicting grades has little social significance" (p. 134). Holland and Richards (1965) have pointed out that a student's extracurricular achievement may be more similar to achievement after graduation than is the academic achievement represented by grades. They demonstrated that academic and extracurricular achievement in college are not highly correlated and urged greater use of extracurricular achievement as an alternative criterion to grades in the development of selection devices. Hoyt (1965) concluded from a review of relevant studies that there is little relationship between college grades and post-college achievement.

The present study grew out of these two sources of dissatisfaction with current methods of predicting success in college. The study has two main goals: First, to find nonintellective predictors that will make effective discriminations between students who have already been highly selected on test score; and second, to find predictors of extracurricular achievement which will, hopefully, be independent of the predictors of grades.

Method

Questionnaire materials were mailed to a sample of 1843 National Merit Finalists and National Merit Scholars in the spring of 1962, shortly before their expected graduation from high school, and also to a sample of 1383 students chosen at random from all students who took the National Merit Scholarship Qualifying Test (NMSQT) in 1961—the same time that this test was taken by the sample of Merit Finalists and Scholars. Usable returns were obtained from 86 per cent of the Merit Finalists and Scholars (the Merit sample) and from 64 per cent of the random sample of NMSQT participants (the normative sample).

A second questionaire was sent to the respondents to the first survey in the summer of the following year, when the students in the sample who attended college should have finished their freshman year. This questionnaire inquired about their college grades, their extracurricular achievement and some additional items for another study. Returns were obtained from 92 per cent of the Merit sample and from 84 per cent of the normative sample. Those who did not attend college for the entire year and those with missing data were discarded leaving a Merit sample of 1330 and a normative sample of 419 for the study.

Predictors

The first questionnaire obtained student responses to four groups of personality, attitude, interest and behavior items: (a) the California Psychological Inventory, a well known personality inventory consisting of 480 True-False items (Gough, 1957), (b) the Vocational Preference Inventory (VPI), an interest inventory consisting of 160 occupational titles to which the student responds Like-Dislike (Holland, 1958, 1965), (c) the Adjective Check List (ACL), adopted with some additions and deletions from a checklist developed by Gough (1960) and in which the student checks from a list of 159 adjectives those he considers descriptive of himself, and (d) the Objective Behavior Inventory (OBI), developed for this study and consisting of 326 things that a student might do (hobbies, sports, leisure time activities, interaction with others, etc.). The student indicates whether he has engaged in each activity frequently, occasionally, or not at all during the past year. In con-

trast to the other three inventories, which inquire about attitudes, opinions, and feelings that are known only to the student himself, the OBI items refer to overt behavioral acts which could be seen by an external observer.

Criteria

Two criteria of college performance were used—academic achievement and extracurricular achievement: (a) Academic achievement was assessed by the student's self report of his freshman grades on a letter grade scale (A, A—, B+, B, B—, etc.). Nichols and Holland (1963) compared such self reported grades with grades calculated from a transcript and found the self reports to be quite accurate ($r = .96$). Davidsen (1963) found similar validity ($r = .92$) for self-reported high school grades obtained in a selection context.

The students attended a variety of colleges, and it seems likely that a given grade may mean quite different things in different colleges or even in different departments within a single university. To the extent that this is the case the grade criterion will be less predictable from student characteristics assessed before college entry. Thus, we would not expect to find validity coefficients as high as might be obtained when predicting them at a specific college. However, a sample of students attending many colleges offers two specific advantages: only predictors that have validity in a variety of college environments will be found; and the validity coefficients will give an indication of the practical utility of the predictors in a setting where predictions must be made before the student's college is known.

The main danger in combining students from different colleges is that the student's grades may be determined in part by the quality of competition from other students at the college. Thus, the analysis may turn up correlates of attending selective colleges rather than of grade getting behavior *per se*. To check this possibility an *ad hoc* sample of 2,000 students attending 246 colleges was drawn from a much larger number available from another study (Astin, 1965; Werts, 1966). The partial correlation of the student's freshman grade average with the selectivity of the college he was attending was —.08 with the student's high school grades partialled out, —.12 with the student's NMSQT score partialled out, and —.18 with both high school grades and NMSQT score held constant statistically.

The index of selectivity was the number of National Merit Commended students choosing the college divided by the number of freshmen admitted (Astin, 1965). The possibility of adjusting college grades for college selectivity was considered, but was not done in this study because the variance in selectivity was not great among the colleges attended by the Merit sample, the relationship of selectivity to grades was not large, and there is no really adequate method of equating grades across colleges. (b) Extracurricular achievement was assessed from the student's report of his nonacademic activities during the freshman year. The questionnaire asked the student to describe all achievements during the past year in the following areas: Leadership, Science, Art, Music, Writing, Speaking, Dramatics, and Athletics. Examples were given of possible achievements in each area and space was provided for the student to write in his achievements. This free response procedure is felt to be an improvement over the checklists used in previous studies (Nichols and Holland, 1963; Holland and Nichols, 1964)—the possibility of error from random and erroneous checking of checklists of very rare events was reduced, and many achievements were obtained that had not been included on the checklists.

The responses in each area of achievement were coded on a three point scale: (1) no achievement, (2) achievement not involving outside recognition, (3) outstanding achievement receiving outside recognition of quality. These achievements were combined into a single three point extracurricular criterion scale by counting those students with no achievement in any area as nonachievers (scored 1), those with one or more outstanding achievements as achievers (scored 3), and the rest in between (scored 2). The few students with recognized leadership achievement, but no other achievement, were scored 2 instead of 3 to keep the criterion scale from being too heavily weighted with leadership. The extracurricular achievements were combined into a single scale because there were not enough outstanding achievers in the individual areas and the lower level achievements were not of sufficient social significance to warrant separate study. However, combining the achievements is justifiable on both empirical and logical grounds: achievements in the various areas tend to have low positive correlations with each other. Since time limitations make it difficult for a student to achieve in more

than one area, the low positive correlations seem to indicate a substantial general tendency to achieve. Roberts (1965) developed scales to predict specific extracurricular achievements and found similar content among the scales. His scales developed to predict achievement in one area also predicted achievement in other areas, and his achievement scales had much higher intercorrelations among themselves than did the actual achievements. Potential users of predictors of extracurricular achievement are more concerned with a general tendency to achieve than with the specific area of achievement, because, like grades, extracurricular achievements in college are of value mainly as a demonstration of a tendency to achieve which is expected to persist to some degree after college.

Analyses

Scales were derived by item analysis against grades and extracurricular achievement using a portion of the Merit sample. After excluding students who did not attend college for a full year and those with incomplete data, 1013 Merit students (those with ID numbers ending with a digit less than seven) were assigned to the derivation group and the remaining 419 to the cross-validation group. No special separation of the sexes was made.

For item analysis against the grade criterion high and low groups of approximately the upper and lower 27 per cent on first year college grades were used: the 262 students who reported a first year grade average of A or A— and also reported receiving some recognition for academic achievement (dean's list, honor society, etc.) composed the high group. The low group consisted of the 246 students with a first year grade average of B— or lower and who received no recognition for academic achievement. For item analysis against the extracurricular achievement criterion, the high group consisted of the 229 students who received outside recognition for one or more extracurricular achievements and the low group consisted of the 294 nonachievers.

Phi coefficients were computed for each of the 1125 items against each criterion. The obi items, which used a three alternatives response format, were dichotomized by combining the middle category with the smallest extreme category. Separate academic and

extracurricular achievement scales for each of the four item formats were formed by combining all items significant at the .05 level. The validity of these scales and their usefulness in combination with test scores and high school rank were checked using the cross-validation group of Merit students and the normative sample.

Results

The characteristics of the scales derived from the four item pools[2] are shown in Table 1. Of the major item types the OBI and the ACL had the largest proportion of significant items and the CPI the smallest proportion. The OBI scales had lower internal consistency than the scales from the other three item pools. These data are not sufficient evidence to establish the superiority of one item pool over another. The validity coefficients of scales composed from the various item pools are the most important consideration, and these are discussed later.

Item Content

The item content of the scales gives an indication of the kind of person who is likely to achieve; although it is important to remember that the items are self reports and that they were identified by group comparisons so that all items do not necessarily apply to all students. With these qualifications in mind the items significantly related to each criterion were grouped into what seemed to be homogenous content categories as follows:

Content of the grade scales. The high grade group, when contrasted with the low grade group, more frequently endorsed items suggesting that they are religious and involved in church activities; are interested in school and value school work highly; are interested in music and participate in musical activities; are hard workers with well established work habits; and are shy, socially withdrawn, and introverted. The low grade group on the other hand more frequently endorsed items suggesting that they are somewhat mischievous and thrill-seeking; are likely to engage in a long list of activities usually considered "fun" for young people; are erratic and undependable; and are interested in occupations which might be considered exciting or dangerous.

[2] The items and scoring keys for the scales developed in this study are available to responsible persons from the author on request.

TABLE 1

Characteristics of Scales Derived from Four Item Pools

Item Pool	Number of items in pool	Grade Scale: Signif. items (p. < .05) *N*	Grade Scale: Signif. items (p. < .05) %	Grade Scale: *Mean*[a]	Grade Scale: *SD*	Grade Scale: *KR-21*	Achievement Scale: Signif. items (p. < .05) *N*	Achievement Scale: Signif. items (p. < .05) %	Achievement Scale: *Mean*[a]	Achievement Scale: *SD*	Achievement Scale: *KR-21*
Objective Behavior Inventory (OBI)	326	107	32.8	63.5	11.2	.80	107	32.8	50.7	11.1	.79
Adjective Check List (ACL)	159	50	31.4	31.2	8.1	.84	54	34.0	25.2	9.2	.86
Vocational Preference Inventory (VPI)	160	35	21.9	25.3	5.8	.82	52	32.5	22.3	9.6	.88
California Psychological Inventory (CPI)	480	110	22.9	69.7	12.1	.83	97	20.8	57.6	12.4	.86

[a] Means, standard deviations and reliability coefficients were calculated from Merit cross-validation sample, both sexes combined ($N = 302$).

The following adjectives were selected as representative of the ACL items significantly differentiating the high and low grade groups: High grade students described themselves as ambitious, capable, conscientious, dependable, efficient, helpful, methodical, modest, patient, quiet, resourceful, self-confident, timid, well-adjusted, and withdrawn. Low grade students described themselves as boastful, carefree, careless, cynical, disorderly, high-strung, impulsive, irresponsible, lazy, messy, rebellious, and sophisticated.

A few individual CPI items deserve mention because they reinforce the general impression from the foregoing discussion that the student who gets good grades is likely to be compulsive and conforming. For example, the high grade group more frequently responded "True" to the following items: "I am stricter about right and wrong than most people." "I would disapprove of anyone's drinking to the point of intoxication at a party." "I keep out of trouble at all costs." "I consider a matter from every standpoint before making a decision." "I always like to keep my things neat and tidy and in good order."

Content of the extracurricular achievement scales. The group of achievers, when contrasted with the nonachievers, more frequently endorsed items suggesting that they engage in a variety of extracurricular activities including music, speech, drama, science, art, etc. (all activities similar to the criterion achievements); are religious and involved in church activities; are outgoing and dominant in interpersonal situations; date frequently; are ambitious and hard working; and are interested in a variety of artistic and intellectual occupations. The nonachievers on the other hand more frequently endorsed items suggesting that they have a low energy level and are overly sensitive to the opinions of others.

The following adjectives were selected as representative of the ACL items significantly differentiating the achieving and nonachieving groups: Achievers described themselves as aggressive, alert, ambitious, artistic, attractive, clever, confident, cooperative, deliberate, dominant, egotistical, energetic, generous, helpful, independent, ingenious, mature, original, persistent, responsible, sophisticated, unconventional, versatile, well thought of, and witty. Nonachievers described themselves as lazy, quiet, shy, slow, and unambitious.

Cross-validation

The intercorrelations of the various scales, test scores, and high school rank (HSR) are shown for the Merit cross-validation group in Table 2 and for the normative group in Table 3.

The correlations shown in these two tables lead to the following conclusions:

(a) The two criteria had low positive correlations with each other for both sexes and both samples. These correlations are low enough for the grade and extracurricular criteria to be considered as independent for all practical purposes.

(b) Among the nonintellective predictors, the CPI scale was the best predictor of grades (except for Merit girls) and the OBI scale was the best predictor of extracurricular achievement. For both criteria the CPI and OBI scales were substantially better predictors (average validity .23) than were the ACL and VIP scales (average validity .11).

(c) The two criteria were differentially predictable by the nonintellective scales. In only five of 32 instances did a scale fail to correlate higher with the criterion for which it was derived than it did with the other criterion, and none of these five instances involved the CPI or OBI. The scales for the two criteria were negatively correlated in the OBI and VIP and positively correlated in the CPI and ACL.

(d) In the Merit sample the NMSQT and SAT had slight predictive validity for both criteria for boys and no validity for either criterion for girls; however, low correlations are to be expected in this instance because of restriction of range of test scores in the Merit sample. In the normative sample the NMSQT had some validity for both criteria for boys but was related only to the grade criterion for girls.

(e) HSR was related to the grade criterion for both sexes and both samples, but not to the extracurricular achievement criterion. It has been suggested that rank in high school class uncorrected for class size is as effective a predictor of college grades as is percentile rank. This was true for boys in both samples, but not for girls.

(f) Two reports of high school rank were available for the Merit sample, one from the school and one from the student. The correla-

TABLE 2

Intercorrelations of Predictors and Criteria in the Merit Cross-validation Sample
Males Below the Diagonal (N = 179); Females Above the Diagonal (N = 138)

	Criteria		Grade Scales				Extracurric.				NMSQT						SAT		HSR			
																			School		Student	
																	V	M	raw	%	raw	%
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. College Grades		09	15	18	−12	24	06	04	12	12	08	05	01	03	−12	−03	01	−10	08	21	04	21
2. Extracurricular Ach.	23		15	15	13	17	17	03	−06	10	03	08	05	−11	−03	03	00	−06	09	05	07	10
3. OBI Grade Scale	29	06		47	07	63	−35	−07	−04	−04	11	12	−12	−05	−17	00	−09	02	14	15	14	18
4. ACL Grade Scale	17	09	27		11	70	−06	52	00	35	03	24	−18	−12	−18	−02	−22	−10	28	19	32	27
5. VPI Grade Scale	11	01	27	11		06	−07	03	−25	04	−09	−06	−09	−09	04	−12	01	−04	−07	−13	09	−16
6. CPI Grade Scale	34	10	68	58	24		−01	21	04	32	04	08	−15	−13	−17	−10	−10	−14	30	23	32	30
7. OBI Extracurricular Scale	04	24	−40	06	−01	−05		28	28	50	−12	−02	01	−06	−06	−10	−05	−17	14	05	10	10
8. ACL Extracurricular Scale	−01	18	−14	58	−06	23	40		15	57	−12	01	−06	−05	09	−05	04	−09	06	−03	10	04
9. VPI Extracurricular Scale	15	18	−08	14	−03	07	37	33		41	−05	00	−05	−03	−17	−11	−11	−08	−01	10	09	22
10. CPI Extracurricular Scale	14	23	03	44	−04	38	46	71	45		−13	01	01	−10	−10	−12	−13	−11	15	04	18	10
11. NMSQT English	15	13	10	13	13	20	18	06	12	04		−02	−05	00	14	46	08	09	−08	04	−09	00
12. NMSQT Math	09	12	00	−04	−01	00	−01	−01	−07	02	01		−11	08	−23	48	−16	50	16	01	09	−01
13. NMSQT Natural Science	03	07	−05	−11	−07	−17	−08	−09	04	−09	03	17		16	34	45	33	02	01	09	09	02
14. NMSQT Social Science	−01	04	−06	−01	−21	−06	−05	−03	−14	−11	−04	04	30		06	50	27	18	−12	−06	−17	−07
15. NMSQT Word Usage	−01	04	−01	−19	04	−07	−09	−07	16	−13	06	−04	29	07		45	61	01	−03	−04	−05	−07
16. NMSQT Composite	10	16	00	−08	−04	−03	−01	−05	03	−09	41	54	66	50	48		39	41	00	01	−10	−05
17. SAT Verbal	17	02	00	−18	00	−05	−08	−18	06	−14	07	10	38	29	51	49		18	−06	−05	−10	01
18. SAT Math	10	06	05	−06	02	−03	−09	−01	−14	−03	04	53	09	17	06	38	19		01	08	−10	−01
19. Raw HSR, School Report	14	02	09	23	05	25	06	19	10	19	12	15	03	04	−05	12	03	12		71	83	61
20. Percentile HSR, School Report	25	04	17	27	14	22	08	22	13	21	09	22	02	−05	−12	09	01	06	57		53	82
21. Raw HSR, Student Report	27	−04	15	26	09	32	06	16	07	17	12	18	−02	−01	08	09	03	14	82	46		70
22. Percentile HSR, Student Report	39	03	24	24	20	28	06	16	12	20	11	22	−05	−14	−12	04	02	12	49	81	64	
23. Sex[a]	15	01	21	05	24	25	40	−05	28	08	28	−33	04	−20	24	−03	11	−38	15	18	11	17

Note.—Decimals omitted.
[a] Correlations with sex are point biserials, with male = 1 and female = 2, computed with the male and female samples combined.

TABLE 3

Intercorrelations of Predictors and Criteria in the Normative Sample
Males Below the Diagonal (N = 201); Females Above the Diagonal (N = 218)

	Criteria		Grade Scales				Extracurric. Scales				NMSQT						HSR	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. College Grades		17	34	22	13	28	−13	00	00	−06	19	16	25	17	11	23	−07	38
2. Extracurricular Ach.	12		02	−13	03	−09	15	07	07	13	16	02	02	−01	03	04	08	00
3. OBI Grade Scale	15	02		34	21	56	−39	−08	−16	−17	−02	−02	−02	−07	−09	−06	−02	26
4. ACL Grade Scale	06	−04	31		03	54	01	51	05	22	−05	−04	08	07	−08	00	−09	15
5. VPI Grade Scale	10	17	15	−12		10	−20	−10	−46	−21	−04	04	−04	00	−04	−02	−07	12
6. CPI Grade Scale	36	02	54	45	10		−06	12	−05	16	−01	01	07	02	−05	01	−07	27
7. OBI Extracurric. Scale	10	30	−42	−07	−07	−08		33	35	55	09	−08	03	01	08	03	13	−06
8. ACL Extracurric. Scale	05	16	−09	51	−16	07	32		24	57	02	−09	04	02	−02	−01	−02	02
9. VPI Extracurric. Scale	22	22	−05	−09	−13	−01	38	19		36	13	−06	20	08	17	13	03	02
10. CPI Extracurric. Scale	18	22	02	17	−08	32	50	48	41		06	−06	08	−02	06	03	02	07
11. NMSQT English	14	13	05	−01	03	−03	08	13	14	10		39	47	44	58	72	20	32
12. NMSQT Math	15	03	−01	05	−02	01	01	14	−02	−01	46		39	44	43	71	15	31
13. NMSQT Natural Science	15	21	−03	−09	19	−09	09	14	12	10	48	41		66	63	81	05	20
14. NMSQT Social Science	09	13	06	03	03	−04	09	16	06	08	60	58	67		59	82	06	25
15. NMSQT Word Usage	21	22	06	−07	18	−04	09	15	20	11	64	52	74	68		82	15	23
16. NMSQT Composite	18	18	03	−02	10	−05	09	18	12	09	77	74	81	82	88		16	34
17. Raw HSR, Stud. Report	31	05	04	−06	04	09	05	−03	11	03	12	07	14	10	15	14		21
18. Percentile HSR, Stud.	34	14	16	03	06	10	05	04	18	06	36	34	39	38	37	45	60	

Note.—Decimals omitted.

tion between these two reports was .81 for boys and .82 for girls. This is a somewhat lower validity for self reported grades than others have reported. A possible reason is that in this study students were asked to report both rank and class size, from which percentile rank was computed. This task may have been more difficult than the usual report of a grade average. The restricted range of class ranks among the Merit students is another possible reason. Whatever the source of error in the self reported HSR may be, it is indeed a peculiar kind of error: the correlation with college grades for girls is the same for self reported HSR as for school reported HSR, but for boys self reported HSR has a substantially, but not significantly, higher correlation with college grades than does school reported HSR.

(g) It is possible to form a rough ordering of the three classes of predictors (test scores, non-intellective scales, and HSR) in terms of their validity for predicting the two criteria, and this order holds in general for both samples and both sexes. To give an indication of the relative size of the coefficients the average validity of the OBI and CPI scales will be used to represent the non-intellective scales, the NMSQT Composite to represent the aptitude tests, and percentile rank to represent HSR. For predicting college grades, high school grades were the best predictor (average validity .33), followed by the nonintellective scales (average validity .27), and finally by the aptitude test (average validity .12). For predicting extracurricular achievement the nonintellective scales were the best predictor (average validity .19), followed by high school grades (average validity .10), and the aptitude test (average validity .07). Grades were predicted with higher validities by all three classes of predictors than was extracurricular achievement.

(h) Girls tended to achieve at a higher level than boys in most respects. They made better grades in high school and college; they obtained higher scores in both groups on nonintellective scales; and they obtained higher verbal test scores. There was no sex difference in extracurricular achievement, and boys obtained higher scores on quantitative tests. The achievement of boys was somewhat more predictable than that of girls: the nonintellective scales had higher validities for boys and in the Merit sample the test scores also had higher validities for boys.

Multiple Correlations

A number of multiple regression equations were computed for predicting each criterion from various combinations of predictors in the various samples. Since these equations showed little consistency from one sample to another and since the multiple correlation coefficients are subject to an unknown amount of shrinkage, they will not be reported in detail here. However, a few general patterns emerged consistently from the various analyses: (a) The nonintellective scales derived from the four item pools did not contribute unique information. Once the scale with the highest correlation with the criterion was taken into account, the partial correlations of the other scales with the criterion were nonsignificant. (b) The nonintellective scales added significantly in most instances to the prediction of the grade criterion that was possible with HSR and the SAT or NMSQT Composite. (c) HSR and the test scores in most instances did not add significantly to the prediction of extracurricular achievement that was possible with a single nonintellective scale.

Discussion

The scales to predict grades developed from the OBI and CPI item pools add significantly to the prediction of college grades that is possible from test scores and high school rank. The extracurricular achievement scales seem to be the best available predictors of extracurricular achievement. Moreover, these scales appear to be particularly robust: developed on a sample of very able students of both sexes, they maintained their validity when applied to boys and girls separately and when applied to a sample of students considerably lower on the ability scale than the sample on which they were derived.

These results warrant tentative use and further trial of these scales in actual selection situations.[3] If the selection context does not greatly affect their validity, the scales will provide an increment in the accuracy of prediction of college performance.

[3] A test of the validity of the scales when used in the selection context of the National Achievement Scholarship Program for outstanding Negro students is now underway. The nonintellective scales may offer a special advantage in the NASP where the value of the traditional predictors may be attenuated by cultural factors.

Extracurricular achievement in this study was not as predictable as were grades. Moreover, the predictors of extracurricular achievement were primarily high school activities and extraversion—traits which may not be considered valuable in themselves. This raises the question of what weight should be given in college and scholarship selection programs to predictors of extracurricular achievement. It is conceivable that some admissions officers may wish to select potential extracurricular achievers to promote campus activities or to balance an overly studious student body. However, in most selection programs the justification for selecting potential extracurricular achievers must be based on the assumption that such achievers will make socially valuable achievements after college. Although this assumption seems reasonable enough, there is little good evidence either to support or to refute it.

There is a clear need for better criteria of success in college, and the lack of good criteria is one of the main difficulties in improving predictors. Since the criterion value of any index of college achievement depends in large part on its relationship to post college achievement, studies of the correlation of college behavior with socially significant achievement after college might help to clarify the criterion problem.

A specific finding which deserves comment is the low predictive validity of the aptitude tests in this study. This might be expected in the Merit sample with its restricted range of scores, but it was also true in the normative sample. The correlations between NMSQT Composite and college grades in the normative sample of .18 for boys and .23 for girls are considerably below the correlations in the .50's reported in the NMSQT Manual (Science Research Associates, 1964) for students attending particular colleges. This shrinkage in validity is probably due to the fact that the students in the present samples attended many different colleges, with the high scoring students attending colleges where the academic competition is greater than for the low scoring students. The shrinkage in validity for the test scores seems greater than for HSR (normative sample validities of .34 for boys and .38 for girls). The great differentiation of colleges on tested ability may effectively attenuate the validity of test scores for predicting college grades in a heterogenous group of colleges, leaving the nonintellective factors, on which colleges are not so differentiated, as the best predictors.

This reasoning would seem to indicate that those who deal with students attending a variety of colleges should develop ways of taking the college into account if they wish to use grades as an index of success.

Validity is often not the sole value in a selection program, and the use of nonintellective scales raises some issues which, although always present, are hidden when intellectual predictors are used. Stalnaker (1965) stated the issue well when he said (prophetically), "In a program very much in the public eye, predictive validity alone cannot rule . . . Suppose there should develop sound evidence that among the highly intelligent, the most conforming, compulsive, dependent, unoriginal individuals do best in college. Should we then try to limit our selection to students having these characteristics" (p. 135)? Stalnaker's hypothetical example is only a slight exaggeration of the content of the nonintellective grade scales. The extracurricular scales on the surface would seem to select more the "All American Boy" type, but should a student be awarded a scholarship because of his broad interests and frequent dates?

The personality traits of the selected students become explicit with the use of nonintellective predictors, but selection for personality characteristics is implicit in all selection programs. For example, unpublished comparisons of the Merit and normative samples used in this study reveal that the NMSQT tends to select a socially withdrawn, studious, introversive character. Since the composition of social groups in our society is increasingly determined by centralized selection programs, more attention should be given to the type of person identified by the various selection strategies. The explicit recognition of the role of personality traits that is inherent in the use of nonintellective predictors may help focus attention on this problem.

Summary

Scales for predicting first year college grades and extracurricular achievement were developed by item analysis from each of four item pools (the California Psychological Inventory, CPI; the Vocational Preference Inventory, VPI; and Adjective Check List, ACL; and an experimental Objective Behavior Inventory, OBI) using a sample of 1013 National Merit Finalists. The scales were cross-validated using samples of 179 male and 138 female Merit Final-

ists and 201 male and 218 female students of average ability. The CPI and OBI scales had higher validities than those developed from the ACL and VPI. The best predictor of college grades was rank in high school class (HSR) followed by the nonintellective grade scales and finally by aptitude test scores. The nonintellective scales added to the prediction of grades in a regression equation including HSR and test scores. The best predictors of extracurricular achievement were the prediction of extracurricular achievement in a regression equation including the nonintellective scales.

REFERENCES

Astin, A. W. *Who Goes Where to College?* Chicago: Science Research Associates, 1965.

Davidsen, O. M. Reliability of Self-reported High School Grades. Unpublished research report, American College Testing Program, 1963.

Gough, H. G. *Manual for the California Psychological Inventory.* Palo Alto, Calif.: Consulting Psychologists Press, 1957.

Gough, H. G. The Adjective Check List as a Personality Assessment Technique. *Psychological Reports,* 1960, 6, 107–122.

Holland, J. L. A Personality Inventory Employing Occupational Titles. *Journal of Applied Psychology,* 1958, 42, 336–342.

Holland, J. L. *Manual for the Vocational Preference Inventory,* (sixth revision). Coralville, Iowa: Educational Research Associates, 1965.

Holland, J. L. and Nichols, R. C. Prediction of Academic and Extracurricular Achievement in College. *Journal of Educational Psychology,* 1964, 55, 55–65.

Holland, J. L. and Richards, J. M. Jr. Academic and Nonacademic Achievement: Correlated or Uncorrelated? *Journal of Educational Psychology,* 1965, 56, 165–174.

Hoyt, D. P. The Relationship between College Grades and Adult Achievement. A Review of the Literature. American College Testing Program, *ACT Research Reports,* 1965, No. 7.

Nichols, R. C. and Holland, J. L. Prediction of the First Year College Performance of High Aptitude Students. *Psychological Monographs,* 1963, 77, No. 7 (Whole No. 570).

Roberts, R. J. The Prediction of First Year College Performance of Very Able Students with Empirically Constructed Scales. National Merit Scholarship Corporation, *NMSC Research Reports,* 1965, 1, No. 5.

Stalnaker, J. M. Psychological Tests and Public Responsibility. *America Psychologist,* 1965, 20, 131–135.

Science Research Associates *NMSQT Interpretive Manual.* Chicago: Science Research Associates, 1964.

Werts, C. E. Career Changes in College. National Merit Scholarship Corporation, *NMSC Research Reports,* 1966, 2, No. 7.

Educational and Psychological Measurement
1966, 26, 917–925.

THE DIFFERENTIAL PREDICTION OF COLLEGE GRADES FROM BIOGRAPHIC INFORMATION

PATRICIA W. LUNNEBORG and CLIFFORD E. LUNNEBORG
University of Washington

A present-day concern of researchers into higher education is prediction of academic accomplishment from nonacademic variables. Holland (1961) was interested in establishing a new basis for awarding scholarships to high aptitude students through the use of nonintellectual indices of creative performance. Other investigators such as Astin (1964) and Heilbrun (1965) have focused upon the dropout problem in higher education and the personal and environmental factors associated with failure. French (1963) wanted to compare aptitude, interest, and personality measures as differential predictors of success in major-field areas as a step in developing a multi-factor battery for use in counseling.

This study was both part of the current trend and apart from it in two important respects. Like French's work, it was concerned with differential prediction in course areas; however, the improvement over intellective predictors stemmed not from more testing—of needs, attitudes, personality traits. The goal was rather to identify useful biographic items from data already at hand in admissions or application forms. These forms remain the same despite the fact that student selection is now based primarily on two kinds of information, overall high school GPA and aptitude and achievement test scores. To be able to use background information would save valuable hours of testing and restore purpose to the time and care students take to complete applications in the first place.

The second unique goal is bound up with the philosophic rationale for differential as opposed to absolute prediction. Differential pre-

diction is based on the assumption that most decisions are made between one course of action or study and another, and that the most useful data to individuals facing such choices are predictions based on differences in performance between areas, rather than predictions based upon what is common among criterion areas. Maximum utility to the individual is placed above maximizing predictive efficiency for the group. In line with making a differential prediction program as comprehensive as possible, biographic variables would serve to expand the factorial composition of an academic battery and, in turn, improve prediction in areas of divergent thinking such as art, architecture, and music, as well as facilitate adding technical and business training criteria for community colleges.

Method

Subjects

The sample consisted of the 520 freshmen (295 males and 225 females) at the University of Washington who took the Washington Pre-College (WPC) Test Battery (1965) when they entered college in fall quarter 1965. Inasmuch as almost all students take this state-wide Battery as high school seniors, the present sample was very atypical in this respect, largely due to the fact that 63 per cent were nonresidents compared with only 9 per cent nonresidents in the total freshman class. The sample also included a comparatively large number (12%) who entered with transfer credit (less than 45 hours). However, all were U. S. citizens, and were born in or after 1940.

Predictors

The predictors consisted of age, sex, and (1) six cumulative high school GPA's: English, foreign languages, mathematics, natural sciences, social studies, and full-credit electives, (2) twelve Battery test scores: vocabulary, English usage, spelling, reading speed, reading comprehension, quantitative skills (3 parts), applied mathematics, mathematics achievement, spatial ability, and mechanical reasoning, (3) 71 items from the admissions application to Washington higher institutions, and (4) 62 items from a "Survey of Col-

lege Plans," a half-hour experimental questionnaire administered with the Battery. Anastasi's (1960) Biographical Inventory items were used as a guide in developing the scoring keys for the latter two predictor sources. Questionnaire items included some from Anastasi's Fordham inventory and from Peterson's survey (1964).

Criteria

The WPC Testing Program provides every student entering a Washington college with grade predictions for 42 college subject areas and for the all-college GPA based upon the Battery, the six high school GPA's, age, and sex. For the present sample the following GPA's based on one quarter of university work served as criteria: over all subjects, English composition, mathematics, foreign language, and physical science (chemistry and physics). Other criteria were, change in major at the quarter's end and withdrawal from school.

Procedure

Predictor intercorrelations and predictor-criterion correlations were computed, and served, together with observed distributions of *S*s responding on dichotomized biographic items, as the basis for choosing 32 biographic variables to be studied further, together with the other (20) predictors.

Sequential predictor selection analyses on these 52 variables were conducted for each of the seven criteria using the iterative predictor selection technique of Horst (Horst and Smith, 1950). No limit was set on the number of variables to be included in each best set so that a greater number of potentially useful variables would be identified. The purpose of these selections was, however, to restrict further the pool of predictors for differential prediction and to insure that this pool included variables which in combination were related to the course areas.

The final pool consisted of the 19 predictors which were among the first six selected above for each of the four course areas. The 19 selected variables were then examined for contribution to differential prediction by the Horst technique (1954), and corrected multiple correlations determined for the eight best predictors with the four course grades and overall GPA.

Results

Table 1 presents the results of the sequential predictor selection of the 52 variables in terms of order of selection and standard partial regression weights. Multiple correlations (R_c) given here and throughout this report are corrected coefficients, i.e., reduced to reflect expected between-sample shrinkage due to sample size and number of predictors as prescribed by Snedecor (1946, p. 348).

The most striking aspect to the selections was the frequency with which biographic variables appeared. Looking across the seven criteria at the first six predictors selected, only one high school grade average may be found and only five test scores, in contrast to twenty-two of the 32 biographic variables. The multiple correlations for the five GPA criteria were substantial in size and not unlike those found when these criteria are more reliable and long-term. When the criteria were nonintellective, withdrawal or change in major, the intellective predictors disappeared entirely from the scene.

Of the 32 biographic variables entered into sequential predictor selection, only five were never selected. Although living on campus as opposed to commuting was positively correlated with grades in

TABLE 1

Order of Selection of Predictors and Standard Partial Regression Weights for Seven Criteria of Freshman First Quarter Performance

(N = 520)

Criterion:	Beta weight
Freshman English GPA (13 predictors selected)	
English usage test	38
Philosophy of higher education: Vocational (Q)	−13
Level of intended vocation (A)	−13
Hours of high school study (Q)	12
Reading comprehension test	18
Major: Humanities (A)	14
Quantitative skills test, Part B (Quantitative Judgment)	15
High school student government activity (A)	10
Expects literary activity in college (Q)	12
Later-born (A)	09
Mother deceased (A)	−08
Father's occupation: Business-organizational (A)	−07

Criterion:	Beta weight
Freshman Mathematics GPA (9 predictors selected)	
Mathematics achievement test	37
Age	30
Father college graduate (A)	21
Level of intended vocation (A)	19
Later-born (A)	16
Expects literary activity in college (Q)	19
Church member (A)	17
Father's occupation: Business-organizational (A)	−12
Hours of high school study (Q)	11
R_c = .42	
Freshman Physical Science GPA (16 predictors selected)	
Intended vocation: Technical (A)	−54
Quantitative skills test, Part B	

TABLE 1. (Continued)

Criterion:	Beta Weight
Age	08
$R_c = .67$	
Freshman Foreign Language GPA (10 predictors selected)	
English usage test	29
Hours of college study (Q)	21
Intended vocation: Technical (A)	−29
Mathematics achievement test	33
Father's occupation: Technical (A)	17
High school foreign language GPA	14
Spatial ability test	−17
Mother college graduate (Q)	−15
Later-born (A)	12
Long-standing college plans (Q)	11
$R_c = .54$	
Freshman Cumulative GPA (11 predictors selected)	
Quantitative skills test, Part B (Quantitative Judgment)	44
High school favorite: Foreign language (Q)	11
Vocabulary test	10
Intended vocation: Technical	−27
Age	32
Long-standing college plans (Q)	17
High school mathematics GPA	17
Hours of high school study (Q)	09
Washington State home town (A)	−10
High school student government activity (A)	09
Mother deceased (A)	07
$R_c = .54$	
(Quantitative Judgment)	41
Age	40
Father's occupational level (A)	−23
Father's occupation: Business-organizational (A)	−11
High school student government activity (A)	23
Philosophy of higher education: Vocational (Q)	22
Hours of high school study (Q)	17
Church member (A)	−14
English usage test	22
High school academic honors (A)	−28
High school mathematics GPA	28
Intended vocation: Verbal-cultural (A)	15
Father's occupation: Technical (A)	19
Level of intended vocation (A)	−14
College motive: to learn certain subjects (Q)	−14
$R_c = .73$	
Withdrawal (5 predictors selected)	
Mother deceased (A)	20
Philosophy of higher education: Nonconformist (Q)	24
High school academic honors (A)	−16
Hours of college study (Q)	−11
Father's occupational level (A)	−10
$R_c = .35$	
Change of Major (8 predictors selected)	
Philosophy of higher education: Nonconformist (Q)	28
Church member (A)	−18
Father college graduate (A)	19
Cultural interest scale (16-point) score (Q)	−14
Intended vocation: Verbal-cultural (A)	13
Father's occupation: Business-organizational (A)	10
Level of intended vocation (A)	−11
Washington State home town (A)	09
$R_c = .35$	

Note.—Decimal points omitted. A indicates application blank item; Q indicates questionnaire item. Levels according to Roe (1956) with 6 representing the lowest level of occupation and 1 representing the most prestigious.

concurrence with Astin's findings (1964), it was highly correlated with level of father's occupation and mathematics achievement, and thus went unselected. Number of high school honors and a scholarship based on scholarship were not selected because of high correlations with high school GPA's and test scores. The other two predictors passed over were preference for an academic or professional life, and having an intellectual philosophy of higher education (65 per cent of the sample). Three predictors overlapping between application and questionnaire were only selected from one or the other source.

Only two of Roe's (1956) occupational types were included in the sequential selection analyses, business-organizational and technical, because the other six types each held only 4 per cent to 10 per cent of the fathers. With a larger sample, the less populated occupations might be expected to contribute to prediction as well, e.g., physical science grades correlated .21 with science as father's occupation. This holds true as well for students' intended occupations of which only the technical and verbal-cultural types (law, teaching, journalism) were popular enough to be studied with this sample.

The results of the selection of variables for their contribution to differential prediction are summarized in Table 2. This selection model (Horst, 1954) insured that the first selected variable, in this instance the English usage test, was the one which best predicted differences among the four criteria at hand. The second selected variable, intent to engage in a technical vocation, was that variable among the eighteen remaining which best extended this differential prediction. Successive additions were designed to include, at each stage, that variable which best complemented the predictors already selected. Table 2 presents only the first eight selected together with their regression weights for corrected multiple correlations for the overall first quarter GPA and freshman course areas. These weight determinations, utilizing once again the sequential technique (Horst and Smith, 1950) assigned optimal weights to each of the eight predictors.

Discussion

The present study offers challenging support for the notion that biographic information of the kind easily available from admissions

TABLE 2

Order of Selection and Standard Partial Regression Weights for Predictors Selected for Contribution to Differential Prediction Efficiency (Decimal points omitted)

Predictors in order of selection (i)	Δ_i	ϕ_i	College GPA criteria: English	Mathematics	Foreign language	Physical science	Overall
English usage test	12665	12665	50	−00	25	15	20
Intended vocation: Technical	05270	17935	−17	−17	−30	−60	−29
Philosophy of higher education: Vocational	05535	23470	−13	−04	−08	20	−09
Level of intended vocation	04817	28281	−10	20	11	−08	00
High school student government activity	04751	33032	10	−07	−08	21	08
Age	03773	36805	10	26	09	34	23
Hours of college study	03429	40234	09	−00	21	−02	06
Mathematics achievement test	02911	43145	19	44	35	44	42
Rc			63	24	46	58	49

Note.—ϕ_i is Horst's (1954) index of differential prediction efficiency for the first i variables selected. At each stage that predictor is selected which will make the greatest contribution (Δ_i) to the index. ϕ_i is proportional to the average of the variances of the predicted criteria minus the average of the covariances among the predicted criteria.

applications or brief questionnaires can effectively contribute to the prediction of academic performance. For each of seven criteria of freshman performance, a pool of 32 biographic variables contributed a greater proportion of predictors compared with intellective variables. Then, of the 19 predictors among the first six selected for the four course criteria, only five were of the pre-college, intellective variety—high school foreign language GPA, a verbal and a quantitative aptitude test, and a verbal and a quantitative achievement test. Lastly, that one or more biographic variables contributed to differential prediction is not surprising; indeed, it was in anticipation of factorial dissimilarity that they were explored. But the selection of so many nonintellective variables prior to something as basic to college grades as quantitative skills or achievement was not expected. Subject to replication, it appears that nonintellective variables of these types possess great potential for the expansion of the differential battery because of their factorial complexity; not only were they different from intellective factors

already present, but they were different from one another. Nonintellective contribution to prediction of courses less like the precollege variables than English composition, physical science, and foreign language courses should be even greater than for these three which are among the six most predictable from the established WPC Battery (1965).

Some discussion is in order concerning the consistency of the relationships between nonintellective variables and academic criteria identified in this study with those already reported. The major area of agreement with Astin's (1964) dropout study was the contribution to college success of socioeconomic level as measured by father's and mother's education, and father's occupational type and level. Even birth order was a useful predictor in the present study, in contrast to the lack of relationship between demographic predictors to grades when the range of talent is restricted (Holland, 1961). Several items which Anastasi (1960) found associated with college success were also confirmed: hours of high school study, vocational goal (Teaching or, in this instance, "verbal-cultural"), high school student government activity, expectation of literary activity in college, language as favorite high school subject, and academic honor(s) in high school. Despite the fact that 76 per cent of the present sample had participated in three or more high school activities, and that 64 per cent anticipated this same level of extracurricular involvement in college, activities were not predictive, while they were in the Fordham study. Two other useful items at Fordham that surprisingly showed no relationship to grades in the present investigation were plans for post-baccalaureate study and anticipation of difficulties concentrating or studying. Study difficulty was correlated solely with withdrawal.

Perhaps the first investigation of biographical factors and academic achievement, Myers' (1952) study at an eastern women's liberal arts college identified college success with the Jewish religion, urban living, foreign born parents, and number of high school offices held, none of which bore any relationship to freshman grades at the University of Washington. These findings, together with the lack of validity for age, father's occupation, and parents' education in Myers' study, recall the substantial restrictions on the present sample which limit any generalizing about the effects of social class or study habits or educational philosophy on college

achievement. Aside from the real differences between the population sampled in these two studies, the relatively large number of variables investigated here in relation to the number of subjects means that in subsequent samples not only will the weights in multiple prediction change, but the contributing correlation coefficients themselves will vary. The importance of the present study lies rather in the confirmation provided French's (1963) conclusion that while intellective measures contribute most to the absolute prediction of college grades, certain nonintellective measures contribute most to differential prediction and thus have a place in prediction when the goal is maximum utility to the individual.

REFERENCES

Anastasi, Anne. The Validation of a Biographical Inventory as a Predictor of College Success. College Entrance Examination Board Research Monograph, No. 1. New York: College Entrance Examination Board, 1960.

Astin, A. W. Personal and Environmental Factors Associated with College Dropouts among High Aptitude Students. *Journal of Educational Psychology*, 1964, 55, 219–227.

French, J. W. Comparative Prediction of College Major-Field Grades by Pure-factor Aptitude, Interest, and Personality Measures. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 767–774.

Heilbrun, A. B. Personality Factors in College Dropout. *Journal of Applied Psychology*, 1965, 49, 1–7.

Holland, J. L. Creative and Academic Performance among Talented Adolescents. *Journal of Educational Psychology*, 1961, 52, 136–147.

Horst, P. A Technique for the Development of a Differential Prediction Battery. *Psychological Monographs*, 1954, 68, No. 9, Whole No. 380.

Horst, P. and Smith S. The Discrimination of Two Racial Samples. *Psychometrika*, 1950, 15, 271–289.

Myers, R. C. Biographical Factors and Academic Achievement: An Experimental Investigation. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1952, 12, 415–426.

Peterson, R. E. Some Biographical and Attitudinal Characteristics of Entering College Freshmen. ETS Research Bulletin, Educational Testing Service, Princeton, New Jersey, 1964.

Roe, Anne. *The Psychology of Occupations*, New York: John Wiley & Sons, 1956.

Snedecor, G. W. *Statistical Methods*, (4th ed.) Ames, Iowa: Iowa State College Press, 1946.

Washington Pre-College Testing Program. *Counselor's Manual 1965–1966*. Seattle: Washington Pre-College Testing Program, 1965.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 927–933.

VALIDITY OF SOME OBJECTIVE SCALES OF MOTIVATION FOR PREDICTING ACADEMIC ACHIEVEMENT[1]

EDWARD J. FURST
University of Arkansas

THE main purpose of this study was to test the hypothesis that a self-rating scale, designed to yield a fairly direct measure of academic achievement-motivation, will show higher correlations with achievement than will a measure of a person's generalized need for achievement. This hypothesis was suggested by Myers (1965) but his study did not present a test of it with data on the same individuals.

Other purposes were to check on the comparative predictive validity of (1) similar scales such as those derived from the Stern Activities Index (SAI), (2) a simple preference between academic and nonacademic subjects (Mayhew, 1965), (3) conventional verbal and nonverbal measures of general ability, and (4) selected combinations of these variables.

Description of Predictor Variables

The following were the predictor variables used:

1. Total score on Achievement Motivation scale
2. Achievement scale (*Ach*), SAI
3. Counteraction scale (*Ctr*), SAI
4. (*Ach* + *Ctr*), SAI

[1] This study represents a part of Project HS-082 supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health. Education, and Welfare, and conducted at the Ohio State University.

Special thanks are due Dr. Albert E. Myers for permission to adapt his version of an achievement motivation scale developed at Educational Testing Service, and Lee C. Lee for help with data processing.

5. Humanities-Social Sciences scale, (*Hum*), SAI
6. Understanding scale (*Und*), SAI
7. (*Hum* + *Und*), SAI
8. Like academic subjects vs. like nonacademic subjects; scored 1, 0
9. Score on verbal measure of general ability
10. Score on nonverbal measure of general ability

The Achievement Motivation scale was an adaptation of one used by Myers (1965). It consisted of the following nine statements to be rated on a five-point scale: 5–very true of self; 4–fairly true; 3–a little true; 2–hardly true; 1–not at all true.

1. When you know there are going to be one or two questions on an examination from outside reading assignments, you always read *all* the material.
2. You regard yourself as a more consistent and harder worker in your classroom assignments than the typical student in your classes.
3. Others (not your good friends) have thought of you as one who "missed some of the fun" because you were so serious.
4. You think your fellow students in school think of you as a hard worker.
5. Most of your teachers probably think of you as one of their hardest workers even though not necessarily one of the smartest.
6. Other interests (sports, extracurricular activities, or hobbies) prevent you from obtaining an excellent rating or mark for *effort* in school work.
7. You have a very strong desire to excel academically.
8. You try harder to get on the school honor roll or merit list than the average student in your class.
9. You try to do most jobs at least a little better than what you think is expected.

Item 6 was scored in reverse. The possible range of scores was from 9 to 45. Retest reliability for a one-month interval was .77 (N = 52). This scale, the simple preference, and the SAI were given early in September, 1965.

Measures of ability differed from sample to sample, depending upon the school's testing program. For the samples as numbered here, the respective verbal measures were: I—California Language

Test, Form X, 1957 Revision, Fundamentals; II—Differential Aptitude Tests (DAT), Verbal Reasoning; III—California Test of Mental Maturity (CTMM), 1957 S-Form, Language; IV—Lorge-Thorndike Intelligence Tests, (LTIT) Verbal; V—CTMM, 1963 S-Form, Language. The respective nonverbal measures were: I—none available; II—DAT, Numerical Ability; III—CTMM, Non-Language; IV—LTIT, Non-Verbal; V—CTMM, Non-Language. Generally, these tests had been given between the middle of the eighth grade and the middle of the ninth.

Samples

The samples consisted of ten classes beginning the ninth grade in September, 1965 and taking a one-semester trial course in economics. Classes taught by the same teacher, and by teachers in the same or closely similar schools and communities, were combined so as to yield groups with larger *N*'s. Each of the five groups consisted of two such classes. The groups varied in ability, as estimated from tables of norms giving centile equivalents for average verbal scores. Groups I and IV were about average—60th and 55th, respectively; III was somewhat above average—76th; II and V were much above average—92nd and 85th—and were largely college-oriented samples from well-to-do suburbs.

Criterion Measures

Final course marks for the two semesters of the eighth grade and the first semester of the ninth comprised the criteria of achievement. So as to make records comparable from sample to sample, only marks in English, social studies, mathematics, and science were used. Letter marks were converted to the usual 0–4 equivalents. Inasmuch as the full eighth-grade record correlated about .8 with the ninth-grade GPA, and inasmuch as these respective GPA's showed a highly similar pattern of correlations with the predictors, the total eighth and ninth-grade GPA was used as the single criterion.

Results

Table 1 is of interest for any evidence of differences between boys and girls. With the exception of the *Hum* scale, boys and girls did not differ significantly on any of the variables. It was therefore

TABLE 1

Means, Standard Deviations, Significance of Differences, and Intercorrelations of Variables for Boys and Girls Separately[a]

Variable	Boys $N = 136$		Girls $N = 92$												
	Mean	S.D.	Mean	S.D.	$M_1 - M_2$	t	1	2	3	4	5	6	7	8	11
1. Ach. Mot.	29.04	5.14	29.23	5.56	−.19	.27		40	19	35	26	21	28	40	53
2. *Ach*	5.22	2.37	4.88	2.22	.34	1.10	40		45	87	27	27	31	24	10
3. *Ctr*	5.86	2.24	5.91	2.02	−.05	.17	27	49		84	30	21	30	14	27
4. *Ach* + *Ctr*	11.08	3.98	10.79	3.61	.29	.56	29	87	86		33	28	36	23	22
5. *Hum*	3.21	2.47	4.11	2.80	−.90	2.54*	40	37	28	38		48	88	29	25
6. *Und*	4.88	2.70	4.67	2.52	.21	.58	45	46	26	42	61		84	34	33
7. *Hum* + *Und*	8.15	4.61	8.78	4.57	−.63	1.03	45	47	31	45	88	89		36	33
8. Acad. Interests	.71	.45	.75	.44	−.04	.67	26	15	19	20	20	29	29		39
11. *GPA*	2.56	.81	2.72	.82	−.16	1.45	47	34	16	29	25	36	33	22	

* Significant at .05 level.

[a] Correlations for boys are below diagonal; for girls, above. Decimal points have been omitted. For boys, $r_{.05} = .17$; $r_{.01} = .22$; $r_{.001} = .28$. For girls, $r_{.05} = .21$; $r_{.01} = .27$; $r_{.001} = .35$.

decided to leave each sample as it was rather than to conduct further studies of boys and girls separately.

Table 2 shows zero-order correlations for the five samples. On the whole (taking the median values), the results confirm the hypothesis that an objective, fairly direct measure of desire to achieve in school is a better predictor of academic achievement than is a measure of generalized need for achievement (*Ach* or *Ach* + *Ctr*), or measures of intellectual interests in general (*Hum* or *Und*, or *Hum* + *Und*, or liking for academic subjects). These correlations agree closely with the corresponding ones in Table 1 for all boys and all girls, respectively. The median correlation of .50 between the Achievement Motivation Scale and GPA is the same as the figure reported by Myers (1965) for his samples of eleventh-grade, college-oriented, students.

The combinations of variables in Table 3 are limited in number because they were chosen so as to bring out certain critical comparisons, rather than to determine the best combination or to study the incremental validity of each predictor. A comparison of the

TABLE 2

Correlations between Predictors and Criterion; Means and Standard Deviations for Measures of Ability[a]

Predictors		Group					
		I ($N = 41$)	II ($N = 41$)	III ($N = 49$)	IV ($N = 37$)	V ($N = 60$)	Md
1. Ach. Mot. Scale		.52***	.44**	.76***	.30	.50***	.50
2. *Ach*		.40*	.00	.13	.33*	.42**	.33
3. *Ctr*		.35*	.11	.27	−.19	.32*	.27
4. *Ach* + *Ctr*		.45**	.07	.24	.09	.40**	.24
5. *Hum*		.12	.19	.29*	.19	.36**	.19
6. *Und*		.17	.02	.30*	.25	.61***	.25
7. *Hum* + *Und*		.18	.12	.36*	.24	.50***	.24
8. Acad. Interests		.28	.11	.46**	.04	.33*	.28
9. Verb. Measure (*V*)		.80***	.28	.48***	.64***	.68***	.64
	$\bar{X}$	10.15	28.56	57.06	50.32	41.90	—
	SD	1.73	7.92	8.93	12.12	8.80	—
10. Non-Verb. Measure (*NL*)		—	.61***	.39**	.27	.56**	.48
	$\bar{X}$	—	20.12	53.98	42.11	37.13	—
	SD	—	7.79	9.10	10.19	6.93	—

* Significant at the .05 level.
** Significant at the .01 level.
*** Significant at the .001 level.
[a] $\bar{X}$ and SD are in raw-score units except grade equivalents in I and T-scores in III.

TABLE 3

Multiple Correlations of Certain Combinations of Predictors and the Criterion

Predictors		Group I ($N = 41$)	Group II ($N = 41$)	Group III ($N = 49$)	Group IV ($N = 37$)	Group V ($N = 60$)
1, 4	R	.58	.44	.76	.23	.54
	F	9.42**	4.62*	31.26**	.96	11.83**
	d.f.	2,38	2,38	2,46	2,34	2,57
7, 8	R	.31	.14	.51	.24	.53
	F	2.07	.39	7.88**	1.08	11.31**
	d.f.	2,38	2,38	2,46	2,34	2,57
1, 4, 7, 8	R	.61	.46	.77	.33	.65
	F	5.45**	2.42	15.97**	.98	9.94**
	d.f.	4,36	4,36	4,44	4,32	4,55
1, 4, 7, 8 9, 10	R	.86	.76	.80	.66	.81
	F	19.34**	7.53**	12.69**	3.82**	17.07**
	d.f.	5,35	6,34	6,42	6,30	6,53

* Significant at the .05 level.
** Significant at the .01 level.

first row of R's with the first row of r's in Table 2 shows that the composite score, $Ach + Ctr$, adds little, if any, to the predictive power of the Achievement Motivation scale. Further, the substitution of variables 7, $Hum + Und$, and 8, liking academic subjects, for variables 1 and 4, results in considerably less predictive power in three of the five samples. This comparison is of importance because it pits measures of sheer desire to achieve against measures of interest in particular intellectual activities. (In factorial analyses of the SAI scales by the writer, these respective pairs of scales came out as independent factors for both the boys and the girls, N's $= 229$ and 165, respectively.) The third row of R's in Table 3 gives an approximation to the probable upper limit of predictive power of the several motivational and interest variables taken as a weighted composite. It may also be noted that the increment over the composite of 1 and 4 was slight in four of the five samples. The last row of R's shows that the addition of one or two measures of general ability to the four-variable motivational-interest composite increased R substantially except in the one sample where R was already on the order of .8. Thus, the motivational-interest composite by itself is not enough for best prediction.

Finally, the data in Table 4 indicate that achievement motiva-

TABLE 4

Correlations of Scores on Achievement Motivation Scale and Measures of General Ability

Measure of General Ability	Group I ($N = 41$)	II ($N = 41$)	III ($N = 49$)	IV ($N = 37$)	V ($N = 60$)
Verbal or Language	.32*	−.12	.38*	.34*	.32*
Non-V. or Non-L.	—	.09	.26	.13	.13

* Significant at the .05 level.

tion is largely independent of general ability, particularly the NV or NL measures, in groups with this level and range of ability.

Discussion

It has been shown that a simple, objective, and fairly direct self-rating scale of motivation to do well in school tends to give better predictions than more generalized measures of need to achieve or measures of intrinsic intellectual interests. The scale essentially samples aspects of the self-concept, reflecting both the person as he sees himself and as he thinks his peers and teachers see him. The writer would interpret the relative superiority of the scale as a special instance of a more general principle—namely, that better prediction results when the elements in the predictor represent as directly as possible the critical elements in the criterion series.

REFERENCES

Mayhew, L. B. Non-Test Predictors of Academic Achievement. Educational and Psychological Measurement, 1965, 25, 39–46.

Myers, A. E. Risk Taking and Academic Success and Their Relation to an Objective Measure of Achievement Motivation. Educational and Psychological Measurement, 1965, 25, 355–363.

Educational and Psychological Measurement
1966, 26, 935–944.

THE PREDICTION OF DIFFERENT CRITERIA OF LAW SCHOOL PERFORMANCE

CLIFFORD E. LUNNEBORG and PATRICIA W. LUNNEBORG
University of Washington

This study was undertaken to identify (1) those pre-admission characteristics of law school applicants which are usefully related to law school performance, and (2) those indices of law school performance which are most predictable. Current studies of law school success characteristically employ the Law School Admission Test (LSAT) and undergraduate grade point average (GPA) as predictors. Typical results of such studies are those of Lewis, Braskamp, and Statler (1964) who reported a correlation of .43 between the best weighting of LSAT and GPA with first semester law grades for 180 students. Of initial concern in this investigation was the determination of other pre-admission factors which might increase the predictability of these commonly used indices. Further, as Locke (1963) has demonstrated, different criteria may be factorially independent, with the result that each of several criteria may be differentially related to the members of a set of predictors. A second concern, then, was whether several criteria of success in legal education were associated with the same or different pre-admissions measures.

Method

Subjects

The complete sample consisted of the 980 students who entered the University of Washington School of Law (UW) from autumn 1956 through autumn 1964. This group was predominantly male (96%), single (61%), and had been graduated from Washington State high schools (78%).

Predictors and Criteria

All data were taken from the individual files maintained on students by the school of law. The 51 predictors listed later were derived from the undergraduate grade transcript, LSAT score report, and law school application form. Undergraduate GPA improvement was defined by subtracting undergraduate GPA from senior year GPA. Father's occupational level and group were coded according to Roe (1956). Locations of home, high school, and college were coded by state to reflect distance from UW. Sixteen criteria (also listed later) were taken from the law school transcript to span the period of law study and beyond.

Procedure

Intercorrelations among the 51 predictors were computed separately for several groups of students, including those entering between 1956–59, 1956–62, 1956–64. For these same groups, correlations between the 51 predictors and the criteria were computed. These predictor intercorrelations and predictor-criterion correlations formed the basis for a series of multiple regression analyses using the iterative selection technique of Horst (Horst and Smith, 1950).

The results reported in this study tend to concentrate upon the 1956–59 group. Because this group was subject to less stringent selection, it displayed wider variability over the range of pre-admission variables than later groups of entrants, which variability is not unlike that present in current groups of law school *applicants*. The 1956-59 sample thus possessed greater practical predictive significance than later samples.

In an initial series of analyses a best set of eight predictors was selected from among the 51 for each criterion. In the Horst solution the best single predictor is first selected, then a second predictor which, among those remaining, best complements the first, followed by a third which best complements the first pair, etc. As a result of these initial analyses, four criteria, ranging from first year law GPA through passing the bar exam, were selected for further study, and the set of predictors reduced from eight to five, four, and two, which were finally assigned weights for establishing predictive equations.

Results

Table 1 presents for the four selected criteria correlations with the 51 predictor variables. Three of these criteria were selected to provide immediate, intermediate, and long-term indices which could be defined for all law school entrants. First year GPA was selected as the most predictable immediate criterion; being dropped for low scholarship by the law school was a highly predictable intermediate measure of success, and passing the bar exam was the post-law school criterion. The fourth criterion, three year cumulative law GPA, was included as an example of reliable postdiction, i.e., a measure highly related to the predictors but based on a very select subgroup of original entrants. The relationship between this criterion and the pre-admission variables must be interpreted quite differently from the first three.

The single predictors that were most highly correlated with the criteria were LSAT total score and the several indices of undergraduate academic achievement with the exception of grade improvement. Outside the intellectual realm the variables that were most highly correlated with law school success were two reflecting delay in entering law school.

The best sets of eight predictors for the 1956–59 group for all 16 criteria are presented in Table 2 with an indication of the order of selection together with the multiple correlations based on the selected sets. These 16 criterion variables were available for all students who entered between autumn 1956 and autumn 1959 with the following exceptions: second and third year GPA, cumulative GPA, and class standing. These six criteria were defined only for those who undertook second and third year studies, and consequently, the interpretation of their relationships to the predictors, as cautioned above, must take this into account. The multiple correlations (R_c) reported in Table 2 and elsewhere in this report are corrected coefficients (Snedecor, 1946, p. 348), i.e., they have been reduced in size to reflect expected between-sample shrinkage owing to sample size and number of predictor variables.

Over the three law years, the multiple correlations for GPA were not appreciably greater than for the purely ordinal class standing. Second and third year performances, based on progressively more restricted samples, showed little diminution of strength of relation-

TABLE 1

Law School Predictor-Criterion Correlations
(Decimal Points Omitted)

	Selected criteria			
Predictors	First year law GPA	Dropped by law school	Three year law cumulative GPA	Admitted to the Bar
1 Undergraduate GPA	41	−30	33	24
2 Senior GPA	37	−29	28	24
3 Undergraduate GPA improvement	05	−05	−00	06
4 Number times on Dean's List	30	−18	25	16
5 Number times on scholastic probation	−22	26	−14	−21
6 Number credit hours of accounting	09	−10	09	08
7 Accounting GPA[a]	33	−25	24	24
8 Student listed two or more areas excelled in	04	−04	02	02
9 Student listed one or more areas failed	−17	19	−17	−19
10 Entry with B.A. degree	−07	14	−03	−18
11 Attendance at more than one undergraduate college	−08	10	−08	−10
12 Major: Pre-Law (A&S)	04	−08	01	07
13 Major: Economics	05	01	07	−04
14 Major: Political Science	02	−04	−05	14
15 Major: Other A&S fields	−09	11	−01	−17
16 Major: Accounting	00	−02	05	−05
17 Major: Business Administration	01	−04	−06	07
18 Major: Engineering	01	−04	06	02
19 Major: "Other" professional schools (forestry, pharmacy)	01	08	03	−05
20 LSAT total score[a]	43	−31	32	30
21 LSAT writing ability ($N = 341$)	26			
22 LSAT general background ($N = 341$)	15			
23 Location of home	−04	06	00	−05
24 Location of high school	−06	09	−04	−10
25 Location of college	−01	03	03	−06
26 Age at entry to law school	−17	18	−20	−22
27 Interval: High school to B.A. or entry to law school[a]	−17	13	−22	−23
28 Interval: B.A. to law entry[a]	−02	10	−01	−05
29 Sex (male)	−02	−02	−11	04
30 Father living	04	−03	10	08
31 Father an attorney[a]	03	−03	01	01
32 Father some college education	01	01	05	−00
33 Mother some college education	07	−08	08	01
34 At least one attorney relative	01	−01	−07	−01
35 Financial reliance on others during law school	07	−06	−00	11
36 Student has one or more dependents	−00	07	−05	−03
37 Number of children	−02	08	−17	−04
38 Prior military service[a]	−10	10	−14	−10

TABLE 1 (Continued)

Predictors	First year law GPA	Dropped by law school	Three year law cumulative GPA	Admitted to the Bar
	Selected criteria			
39 Intent to practice law	00	−02	−09	05
40 Intent to teach law	−02	05	08	−03
41 Intent to work part-time during law school	−08	11	−09	−15
42 Number hours work/week while in law school	−11	12	−03	−13
43 Father's occupational level (1, professional through 6, unskilled)[a]	−06	02	−06	02
44 Father's occupation: Social/personal service	−00	03	00	−08
45 Father's occupation: Sales	−03	08	06	−05
46 Father's occupation: Business, organizational, governmental	05	−08	−01	09
47 Father's occupation: Technical	−03	−02	−02	01
48 Father's occupation: Outdoors	−01	03	−01	01
49 Father's occupation: Science	04	04	03	−03
50 Father's occupation: Verbal-Cultural	04	−06	−01	04
51 Father's occupation: Performance-Cultural	−05	02	00	−02
Law classes included	1956–64	1956–62	1956–62	1956–59
N	980	750	514	393
N never less than:	885	677	466	320

Note.—Variables 8–19, 29–36, 38–41, and 44–51 are dichotomous.
[a] Indicates some missing data.

ship; indeed, the three year cumulative GPA based only on those subjects completing the three years was as highly related to the predictors as first year GPA. Voluntary withdrawal from law school was weakly related to the predictors and probably shared little variance with other criteria of academic achievement.

Predictor variables most often included among the first four selected, as Table 2 indicates, were LSAT total score, undergraduate GPA, interval: high school to B.A. or law school entry, major: arts and sciences fields other than economics, pre-law or political science, entry with B.A. degree, and father's occupation: business, organizational, governmental.

Eight predictors were selected for further analysis as summarized in Table 3. The above-mentioned six were included with age at entry substituted for the strongly correlated (.69) high school-college completion interval inasmuch as the age measure was the

more easily defined and computed. Two other variables, accounting GPA and student listed one or more areas failed, were included because each was specifically related to two or more of the four selected criteria. For each of the four criteria picked for further analysis the best sets of five, four, and two predictors were selected from among these eight. Standard partial regression weights and corrected multiple correlations were determined in each analysis.

LSAT total score was among the first two predictors selected for each criterion. Undergraduate GPA, interestingly, was chosen when the criterion was also a continuous GPA; when the criterion was dichotomous, a dichotomous measure of undergraduate academic success was chosen instead, i.e., student listed one or more areas failed. Finally, the set of five selected predictors for each of the three predictive criteria was completed by the same three variables, age at entry, accounting GPA, and major: other arts and sciences fields.

Discussion

The most important conclusion from the analyses is that there are several items of biographic or educational information which can be

TABLE 2

Best Sets of Eight Predictors for Sixteen Law School Criteria (N = 393 Entering 1956–59)

Criteria	Multiple Correlation (R_c)	Order of predictor selection[a]
First year GPA	635	20, 2, 27, 51, 15, 42, 32, 7
First year class standing	619	20, 2, 42, 15, 7, 9, 51, 29
Obtained a second year GPA	443	20, 15, 41, 5, 11, 2, 19, 50
Second year GPA	633	1, 20, 8, 46, 29, 19, 39, 24
Two year cumulative GPA	616	1, 20, 46, 19, 24, 8, 13, 7
Second year class standing	586	1, 20, 46, 27, 19, 7, 8, 24
Obtained a third year GPA	463	20, 27, 10, 15, 2, 41, 29, 36
Third year GPA	563	1, 20, 8, 39, 46, 45, 29, 7
Three year cumulative GPA	638	1, 20, 19, 39, 46, 35, 24, 7
Third year class standing	603	1, 20, 19, 24, 39, 28, 7, 34
Law Review member	407	1, 20, 15, 10, 24, 32, 37, 36
Received "honors" as law student	444	1, 20, 27, 36, 46, 47, 42, 33
Admitted to the Bar	468	20, 27, 10, 15, 2, 13, 9, 29
Withdrew from law school	246	41, 29, 16, 35, 45, 25, 4, 11
Dropped by law school	521	20, 2, 26, 9, 45, 11, 5, 29
Graduated from law school	494	20, 27, 10, 5, 41, 15, 45, 36

[a] Predictor identification numbers from Table 1.

TABLE 3

Standard Partial Regression Weights for the Prediction of Four Criteria of Law School Success
(Decimal Points Omitted)

Predictors	First year law GPA			Dropped by law school			Three year cumulative law GPA			Admitted to the Bar		
Number of predictors weighted	2	4	5	2	4	5	2	4	5	2	4	5
LSAT total score	34	34	29	−34	−25	−25	30	29	24	29	29	24
Undergraduate GPA	31	30	25				39	41	36			
Age at entry		−16	−16		18	16				−20	−16	−16
Accounting GPA			18		−20	−21			15			14
Major: other A&S fields[a]		−14	−14			08					−15	−16
Student listed one or more areas failed				22	18	18					−16	−15
Father's occupation:												
Business								11	10			
Entry with B.A.								08	07			
R_e	53	58*	59*	40	47*	47*	56	57	59*	36	41*	43*
N (classes entering 1956–59)		393			393			258			393	

* Multiple correlations significantly larger (.01 level) than correlations based on two predictors.
[a] A&S major other than economics, political science, and pre-law.

used together with the LSAT total score and a measure of overall undergraduate academic performance to increase the predictability of success in law school. For each of the four criteria selected for final study, multiple correlations based upon five variables were significantly greater than those utilizing only the two measures. Furthermore, the additional contributing variables can be the same whether the predictive criterion is immediate, intermediate, or long-term.

Two of the five selected predictors merit additional comment. In a related study (Lunneborg and Lunneborg, 1966), first year law grades were found to be negatively related to those delays in education which occurred prior to attainment of the baccalaureate degree, but were unrelated to those occurring after the baccalaureate. Because age at entry indexes both of these kinds of delay, its predictive value may be attenuated. Its practical value in admissions formulae over calculating periods of delay is hoped to offset this loss. The predictor variable, major: other arts and sciences fields, serves to dichotomize law school entrants into those whose undergraduate program was one of professional preparation (engineering, business administration, forestry, etc.) or preparation for a legal profession (pre-law, economics, political science) and into those who majored in some less professionally-oriented arts and sciences subject such as English, history, or physics. It is tempting to look upon this variable as one which distinguishes those with professional educational goals from those who sought a liberal arts education.

Variables of interest that did not prove to be of predictive value included the LSAT part scores (general background and writing ability) and improvement in undergraduate grades (senior year GPA minus cumulative GPA). Although LSAT part scores could not be included in the 1956–59 analyses reported here, they were part of the pool available for predicting first year law grades for entrants between 1960-64. Neither part score was ever included among the first eight predictors selected. The undergraduate grade improvement measure was developed because of the belief in law school admissions committees that later undergraduate work is more valid than earlier work. It was thus hypothesized that students who showed a negative discrepancy between their senior year and cumulative GPA would perform most unsatisfactorily in law school,

and that those who showed little or no discrepancy would be less apt in law study than the "late bloomers" who should show a large positive difference when their cumulative GPA was subtracted from their senior year GPA. This hypothesis was not supported.

The strong relationship between pre-admission variables and three year cumulative GPA is noteworthy for several reasons. Rather than providing a basis for predicting success in law school for all entrants, it represents an explanation (a postdiction) of the three year success for the survivors in the system. Because of the attrition which occurred prior to completion of the third year, the magnitude of the relationship which exists for this postdictive criterion is quite surprising. Differences among the more select third year students were apparently just as closely related to pre-law variables as the differences among the first year students. Distefano and Bass (1959) observed similarly among a highly restricted sample of practicing lawyers that LSAT and pre-law grades discriminated between those rated as high and low in legal ability.

Summary

A series of multiple regression analyses was undertaken to identify pre-admission variables useful in predicting several criteria of law school success. LSAT total scores and undergraduate academic performance were consistently the strongest predictors. Age at entry, accounting grades, and undergraduate major were shown to increase the predictability not only of first year law school GPA but also of continuance in school and eventual passage of the bar examination.

REFERENCES

Distefano, M. K. and Bass, B. M. Prediction of an Ultimate Criterion of Success as a Lawyer. *Journal of Applied Psychology*, 1959, 43, 40–41.

Horst, P. and Smith, S. The Discrimination of Two Racial Samples. *Psychometrika,* 1950, 15, 271–289.

Lewis, J. W., Braskamp, L., and Statler, C. Predicting Achievement in a College of Law. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 947–949.

Locke, E. A. The Development of Criteria of Student Achievement. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1963, 23, 299–307.

Lunneborg, C. E. and Lunneborg, Patricia W. Relations of Back-

ground Characteristics to Success in the First Year of Law School. *Journal of Legal Education*, 1966, 18, 425–436.
Roe, Anne. *The Psychology of Occupations*. New York: John Wiley & Sons, 1956.
Snedecor, G. W. *Statistical Methods*. (4th ed.) Ames, Iowa: Iowa State College Press, 1946.

Educational and Psychological Measurement
1966, 26, 945–953.

VERIFICATION OF SIX PERSONALITY FACTORS

ANDREW L. COMREY and KAY JAMISON
University of California, Los Angeles

Previous factor analytic personality research papers in this series have been based upon the assumption that it is desirable to factor analyze total scores over collections of relatively homogeneous items, rather than to analyze single items (Comrey, 1961, 1962, 1964, 1965, 1966; Comrey and Schlesinger, 1962). Items have been developed to fit a certain homogeneous item pool, but have been required to exhibit a statistical as well as a logical belonging before being actually assigned to that item group.

On the basis of these previous findings, six major personality factors were selected for further study. Six factored homogeneous item dimensions (FHIDs) were chosen to represent each hypothesized factor. Each FHID, in turn, consisted of six items intended to be relatively homogeneous in content. Most of the FHIDs selected for a given factor had been good marker variables for that factor in previous analyses; some were newly developed for this study. The major purpose of the present investigation was to determine whether the postulated six-factor personality structure would emerge as predicted. There were also two secondary objectives. The first of these was to determine whether the expected factor structure could be obtained despite an important departure from previously used sampling procedures. The second minor objective was to obtain some information about the effect of equating the number of positively and negatively worded items in each FHID. In previous investigations, no systematic attempt was made to control this possible source of bias.

Procedure

Sample

The personality inventory prepared for this study was distributed to an area sample of people in Van Nuys and Santa Monica, California. Both of these areas are middle-class communities with individual sub-areas ranging in economic status from lower to upper middle class. The sample was drawn by numbering all the blocks in the two areas and then choosing ninety numbers at random from a total of 1570 possibilities. Four of these blocks were rejected because they were non-residential. Every home on a selected block was visited to explain the study and to invite participation. To encourage participation, an individual personality analysis was promised to each participant. The sample used consisted of the first 274 answer sheets returned, 153 females, and 121 males. The mean age was 39.4 and the mean number of years of schooling completed was 13.2. The respective standard deviations were 16.3 and 2.5. In previous samples, respondents were not contacted personally but merely given a written invitation to participate. Furthermore, a substantial part of each sample consisted of university students.

In the present sample, a total of 1152 dwelling units was visited; 500 units visited had no one at home; 234 units refused to participate; 418 units produced at least one person willing to participate. A total of 737 inventories was distributed in this manner but only 274 were completed and returned. The most common reasons given for refusal to participate were: not interested, not enough time, too old, too sick, and language problems. Suspicions about the nature of the project, misgivings about who would see the results, antipathy toward psychology, and antipathy toward UCLA were less frequently mentioned reasons.

Inventory

The inventory consisted of 216 items designed to measure six hypothesized factors: Shyness, Dependence, Empathy, Neuroticism, Compulsion, and Hostility. Thirty six items were allocated to each factor, divided among six FHIDs of six items each. For example,

the Dependence factor comprised the following six FHIDs: Succorance, Lack of Self Sufficiency, Deference, Conformity, Need for Approval, and Affiliation. Under each FHID were six items which were expected to correlate substantially with each other. In addition, each FHID designed to measure Dependence was expected to correlate with each other such FHID.

In an attempt to achieve higher item reliability than that typically obtained with forced-choice and true-false item formats, two response scales were used which had nine alternatives. Scale X had the following possible answers: 1. Never, 2. Almost Never, 3. Rarely, 4. Occasionally, 5. Fairly Often, 6. Frequently, 7. Usually, 8. Almost Always, and 9. Always. Scale Y had these possible responses: 1. Absolutely Not, 2. Very Definitely Not, 3. Definitely Not, 4. Probably Not, 5. Possibly, 6. Probably, 7. Definitely, 8. Very Definitely, and 9. Absolutely. The subjects were informed that they could use either scale but next to each item number on the inventory an X or a Y was marked to suggest the scale most probably suitable.

Half the 216 items were worded positively with respect to the FHID name and half were worded negatively. For example, two items on the FHID, Reserve, were, "I do less than my share of the talking in a conversation," and "I do more than my share of the talking in a conversation." These two items were well-separated in the inventory booklet, as were all items measuring the same FHID. An earlier investigation found a measure of acquiescence to have little relationship to these factors (Comrey, 1964). For purposes of further investigation, however, the FHIDs in this study were composed to control this possible source of bias.

The 36 FHIDs used in this investigation are listed below with a sample item from each. A reliability estimate is given in parentheses. This estimate was obtained in each case by averaging the interitem correlations to obtain an estimate of a one-item test and then correcting this figure by the Spearman-Brown formula to a test of the number of items used. Items follwed by an asterisk are negatively worded with respect to the dimension name.

1. *Seclusiveness* (.85) lX. I dislike being with a lot of people I don't know.
2. *Succorance* (.79) 20X. I dislike being dependent on someone else.*

3. *Generosity* (.70) 75X. I am willing to share what I can with others less fortunate.
4. *Inferiority Feelings* (.77) 76X. I feel inferior to the people I know.
5. *Grooming* (.84) 23X. I am unconcerned about how my clothes look to others.*
6. *Hostility* (.78) 24Y. Most people are valuable human beings.*
7. *Reserve* (.86) 79X. I have little to say in a group.
8. *Lack of self-sufficiency* (.86) 8Y. I dislike being left alone.
9. *Service* (.73) 81Y. I would like a job in which I help people who have problems.
10. *Depression* (.82) 82X. I feel that life is drudgery and boredom.
11. *Cautiousness* (.86) 29X. I like to live dangerously.*
12. *Rhathymia* (.64) 84Y. I think it is better to enjoy life than to do something worthwhile for society.
13. *Shyness* (.89) 13X. I find it difficult to talk with a person I have just met.
14. *Deference* (.63) 86Y. I would like a regular job as an assistant to a really great man.
15. *Helpfulness* (.79) 15Y. I enjoy helping people even if I don't know them very well.
16. *Pessimism* (.86) 88Y. I am inclined to be a pessimist.
17. *Order* (.89) 17X. I keep everything in its proper place so I know just where to find it.
18. *Cynicism* (.81) 180Y. Most students in school would rather fail than cheat.*
19. *Stage Fright* (.80) 109Y. It would be hard for me to do anything in front of an audience.
20. *Conformity* (.67) 38X. I feel better doing what everyone else is doing.
21. *Interest in People* (.72) 183Y. I would prefer a job where dealing with people is the most important part of it.
22. *Agitation* (.86) 130X. I relax without difficulty.*
23. *Love of Routine* (.76) 41X. I like to maintain a regular schedule of daily activities.
24. *Psychopathy* (.80) 114Y. If I were in business, I would lie if necessary to make money.
25. *Submission* (.76) 205X. I manage to make my presence felt in a group.*

26. *Need for Approval* (.83) 62X. I ignore what my neighbors might think of me.*
27. *Sympathy* (.77) 45X. I am a very sympathetic person.
28. *Inadequacy* (.78) 208Y. I think I have a lot of ability.*
29. *Meticulousness* (.78) 119X. When I do a job, I try to make it perfect down to the last detail.
30. *Defensiveness* (.71) 66X. When I have to compete with other people, I find them to be friendly and fair.*
31. *Follower Role* (.81) 193X. I refuse to compete in struggles for power.
32. *Affiliation* (.63) 50X. I am loyal to my friends even when they are wrong.
33. *Tolerence* (.65) 123X. I like to associate with people who come from a very different background than my own.
34. *Lack of Ego Strength* (.83) 142X. I feel able to deal with the problems I face.*
35. *Drive to Finish* (.86) 215X. It is difficult for me to keep at something until it is finished.*
36. *Aggression* (.74) 144X. I try to avoid saying anything that might make somebody mad.*

Analysis

Preliminary item analyses were carried out to determine which items should be retained for the final analysis of FHIDs. The items were divided into six groups of 36, with all the items from one FHID in the same group. Each factor had a FHID in every group of 36 items. Each group of 36 items had one and only one set of six items from every hypothesized factor. A factor analysis of items was carried out for each of these six groups of 36 items. In all of these analyses, a factor emerged as expected for every FHID included. Some of the factors were less clear than others, however, and several items had to be dropped because they failed to achieve sufficiently high loadings on the appropriate item factor. Fifteen of the FHIDs had all six items retained, 10 had five items retained, seven had four items, three had three items, and one FHID had only two of the original six items retained. Total FHID scores were computed by adding up the scores on the retained items. Negatively worded items were reversed by subtracting the item score from 10 before adding. The intercorrelation matrix of Pearson co-

efficients was computed for these 36 FHID scores. To check for the presence of alternate forms among the variables, high values in the correlation matrix were corrected for attenuation to determine if they would approach 1.0; in no case did this occur.

The 36 × 36 matrix of uncorrected correlations among FHID scores was factor analyzed by the minimum residual method (Comrey and Ahumada, 1964, 1965). Twelve factors were extracted before encountering convergence on vectors of opposite sign, establishing 12 as the upper limit on the number of valid factors. These 12 factors were rotated orthogonally by the normal criterion I of the Tandem Criteria for analytic rotation in factor analysis (Comrey, 1967). Only six major factors appeared in the criterion I rotations. These six criterion I factors were further rotated by the normal criterion II of the Tandem Criteria. Criterion I attempts to crowd the variance on as few factors as possible, subject to the limitation that variables on the same factor must be correlated. This makes it easier to identify and eliminate minor factors. After eliminating the minor factors, the remaining factors are rotated by criterion II which approximates simple structure where the data permit such a solution. Criterion II accomplishes this by rotating to reduce the loadings of uncorrelated variables on the same factors.[1]

Results

Loadings of .3 or more on the six criterion II factors will be given below. If a variable was hypothesized to appear on the factor, its loading will be given even if it is less than .3. If a variable was not hypothesized to be on the factor, the factor loading will be followed by an asterisk.

I. *Shyness*. 1. Seclusiveness, .63; 7. Reserve, .59; 11. Cautiousness, .35;* 13. Shyness, .62; 19. Stage Fright, .65; 25. Submission, .48; 31. Follower Role, .70; and 36. Aggression, —.41.*

---

[1] The computations for this study were carried out on the IBM 7094 operated by the UCLA Computing Facility. The following materials have been deposited with the ADI Auxiliary Publications Project, Photoduplications Service, Library of Congress, Washington, D. C. 20540: Correlation matrix, minimum residual factor matrix, criterion I rotations, criterion II rotations, test booklet, answer sheet, and list of items for each FHID and factor. A copy may be secured by citing document number 9053 and remitting in advance $1.75 for 35 mm. microfilm or $2.50 for photoprints. Make check payable to: Chief, Photoduplication Service, Library of Congress.

II. *Dependence.* 2. Succorance, .53; 8. Lack of Self Sufficiency, .52; 14. Deference, .28; 20. Conformity, .67; 26. Need for Approval, .53; and 32. Affiliation, .40.

III. *Empathy.* 1. Seclusiveness, —.38;* 3. Generosity, .70; 9. Service, .55; 15. Helpfulness, .74; 21. Interest in People, .65; 27. Sympathy, .70; and 33. Tolerance, .38.

IV. *Neuroticism.* 2. Succorance, .30;* 4. Inferiority Feelings, .64; 10. Depression, .73; 16. Pessimism, .71; 22. Agitation, .66; 25. Submission, .37;* 28. Inadequacy, .79; and 34. Lack of Ego Strength, .77.

V. *Compulsion.* 5. Grooming, .40; 11. Cautiousness, .48; 12. Rhathymia, —.36;* 17. Order, .69; 23. Love of Routine, .53; 29. Meticulousness, .61; and 35. Drive to Finish, .61.

VI. *Hostility.* 6. Hostility, .64; 12. Rhathymia, .50; 18. Cynicism, .73; 24. Psychopathy, .59; 30. Defensiveness, .51; and 36. Aggression, .36.

Six criterion I factors were regarded as too minor to include in the criterion II rotations. Three of these had no loading as high as .3. Loadings of .3 or more will be given below for the remaining three residual factors.

Residual. 11. Cautiousness, .37; and 23. Love of Routine, .32.

Residual. 5. Grooming, .41; and 17. Order, .33.

Residual. 30. Defensiveness, .37.

Discussion

In examining the results, it is clear that six major personality factors did emerge as hypothesized. The remaining six factors were very minor in importance by comparison with the first six. Every FHID expected to appear on a given factor, except Deference on the Dependence factor, did have a loading of at least .3 on that factor. Neuroticism was loaded by all six of its defining FHIDs to the extent of .5 or more; Shyness, Empathy, and Hostility had five out of six FHIDs with loadings of .5 or more; and Compulsion and Dependence both had four of their six FHIDs with loadings of .5 or more. No FHID had a loading in excess of .41 on a factor it was not intended to define. The few loadings of .3 or more for FHIDs on factors they were not intended to define were reasonable. The negative loading of —.41 for Aggression on the Shyness factor

and the loading of −.36 for Rhathymia on the Compulsion factor are both in line with what might be expected.

A statistical comparison of the mean factor scores for males and females in this sample is given in Table 1. These factor scores were obtained by computing the total scores over those items retained for each FHID on the basis of the item analyses. Thus, scores for a given factor were obtained by adding the total scores for those FHIDs designed for that factor. In most cases, the retained items were those which had loadings of at least .45 on the item factor which emerged for that FHID in an analysis of items. The FHIDs Rhathymia and Affiliation, however, did not produce good item factors. For the Rhathymia FHID, four items were selected which had the highest intercorrelations. These intercorrelations ranged from .27 to .35. For the Affiliation FHID, two items were selected for retention which correlated .48 with each other. The matrix of correlations among the six major factor-score variables is given in Table 2.

With reference to the primary objectives of this investigation,

TABLE 1

Means and Standard Deviations for the Factor Scores

	Males		Females		
Factor	*M*	*SD*	*M*	*SD*	*t*
Shyness	82.9	19.7	85.1	22.0	.86
Dependence	99.4	19.4	106.5	21.5	2.82**
Empathy	95.9	16.7	104.4	14.8	4.44**
Neuroticism	61.5	15.9	66.3	20.6	2.11*
Compulsion	141.4	19.9	141.5	21.1	.40
Hostility	83.5	19.1	80.0	20.2	1.45

* Significant at the .05 level.
** Significant at the .01 level.

TABLE 2

Correlations among Factor Scores

	1	2	3	4	5
1					
2	.11				
3	−.24	.28			
4	.35	.30	−.08		
5	.16	.04	.08	−.23	
6	−.11	−.22	−.44	.27	−.32

it seems evident that the postulated six-factor personality structure did emerge as predicted. Since these factors have been identified in several previous studies in this series, it seems reasonable to describe them as well-defined and consistent factored measures of personality. Furthermore, the factor structure emerged despite a substantial departure from previous sampling procedures, suggesting that these factors have some degree of stability over different kinds of subject populations. Finally, the conversion of items to an even split between positively and negatively worded statements resulted in essentially the same factors as those obtained when the statements were all worded in one direction.

REFERENCES

Comrey, A. L. Factored Homogeneous Item Dimensions in Personality Research. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 417–431.

Comrey, A. L. A Study of Thirty-five Personality Dimensions. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1962, 22, 543–552.

Comrey, A. L. Personality Factors Compulsion, Dependence, Hostility, and Neuroticism. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1964, 24, 75–84.

Comery, A. L. Scales for Measuring Compulsion, Hostility, Neuroticism, and Shyness. *Psychological Reports*, 1965, 16, 697–700.

Comrey, A. L. Comparison of Personality and Attitude Variables. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1966, 26, 853–860.

Comrey, A. L. Tandem Criteria for Analytic Rotation in Factor Analysis. *Psychometrika*, 1967, in press.

Comrey, A. L. and Ahumada, A. An Improved Procedure and Program for Minimum Residual Factor Analysis. *Psychological Reports*, 1964, 15, 91–96.

Comrey, A. L. and Ahumada, A. Note and Fortran IV program for Minimum Residual Factor Analysis. *Psychological Reports*, 1965, 17, 446.

Comrey, A. L. and Schlesinger, B. Verification and Extension of a System of Personality Dimensions. *Journal of Applied Psychology*, 1962, 46, 256–262.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 955-963.

THE DEVELOPMENT OF A SCALE TO MEASURE ATTITUDINAL DIMENSIONS OF THE EDUCATIONAL ENVIRONMENT[1]

JOHN K. TUEL
Oral Roberts University
AND
MERVILLE C. SHAW
University of California, Los Angeles

A large scale research project was initiated in 1963 for the purpose of testing a rationale for guidance services which postulated that the prime objective of the guidance specialist should be to effect positive changes in the educational environment (Shaw and Tuel, 1964a). Two sub-objectives of the study were: (1) to assess the nature and attitudinal structure of the educational environment (Tuel, 1964a,b), and (2) to attempt to modify the attitudes of significant persons in the educational environment (Shaw and Tuel, 1964b). Essential to both of these objectives was the utilization of a broad-spectrum attitude inventory capable of administration to various educational role groups including students, parents, teachers, guidance specialists, and administrators. The major purpose of such an inventory would be to reflect basic attitudes towards educational philosophies and practice.

Review of the relevant literature revealed that the topic had received but scant attention from research workers, and that such an inventory meeting the above specifications did not then exist. Necessity then led to the development, described below, of the School Opinion Survey (Tuel and Shaw, 1964), a scale designed to

[1] This research was a facet of a larger project supported by a grant from the National Institute of Mental Health administered through the Interprofessional Research Commission on Pupil Personnel Services.

meet both the specific needs of the current research project and what seemed to be a more general need for an instrument to assess the education-related attitudes of the significant persons in the learning environment.

The School Opinion Survey was built upon several assumptions. The first was that the "educational environment" is a useful construct describing that part of the total effective environment which directly influences academic learning (Tuel, 1964b). The second was that the educational environment exhibits several levels extending from the intrapersonal (Tuel and Wursten, 1965b) through a widening scope to the national level (Tuel and Wursten, 1965a). A third assumption was that philosophy, values, objectives, and techniques form a continuum or hierarchy from the abstract to the concrete, from the philosophical to the operational, or, looking at it still a third way, from the attitudinal to the behaviorial. A fourth postulate was that there is some degree of congruence among the philosophy and values espoused by an individual, the objectives which he sees as being appropriate for education, and the techniques he favors in carrying out educational procedures. Finally, it was proposed that attitudes along this continuum and their degree of congruence could be measured.

Procedure

A pool of over three hundred items reflecting various philosophical positions, values, educational objectives, and techniques for accomplishing these objectives was assembled and tried out on various small samples. Items which proved ambiguous or which showed a restricted range of responses were dropped from this pool. Since the questionnaire was designed for use with the entire gamut of educational role groups, i.e., with parents, teachers, administrators, and counselors at all school levels, as well as with students at the high school level, vocabulary level and meaningfulness had to be carefully controlled. Preliminary screening left a pool of 250 items. These items were then administered to all teachers, principals and counselors, and to all tenth grade students and their parents in a medium sized California city school district. The resulting data were subjected to factor analysis with varimax rotation.

Results and Discussion

Ten prominent factors were extracted, which together accounted for over 70 per cent of the total variance. The first three of these factors appeared to represent the major positions in educational philosophy, as outlined by Brackenbury (1959) and subsequently shown by Tuel (1966) to have been operative in the development of American higher education. The remaining seven factors represented techniques for implementing various patterns of educational philosophies.

The first and strongest factor contained items chiefly concerned with individual differences, individual personal development and human objectives as opposed to strictly intellectual subject-matter-oriented objectives. Educational objectives stressing self-realization of the consequences of one's own behavior, personal enjoyment of learning, development of critical thinking, and breadth of curriculum tended to fall into this factor. All these objectives presupposed a philosophical structure which saw reality as centered in the human being as an individual and a value system focused on the enhancement of the individual personality. This factor has thus been named the *Humanist* factor.

Factor II appeared qualitatively to be the antithesis of Factor I, i.e., a sort of "antihumanist" scale. However the two factors actually proved to be uncorrelated (See Table 1). Factor II stressed intellectual development, attention to "objective facts," avoidance of controversial issues, and the irrelevance to education of individual differences in feelings, interests and motivation. Such values evidently stemmed from a philosophical system which located reality *outside* of the individual in ideas, society, or the material universe, i.e., idealism, social relativism, or realism. Since the external "objective" component seemed strongest, it was decided to designate this the *Realist* factor.

Factor III presented an interesting pattern. The scale was composed of two equal parts of opposite valence. Half of the items took a social relativist or experimentalist position stressing the relativity of knowledge and truth, the irrelevance of religion to education, and the social origin and changeability of moral law. The other half of the items (which were all negatively correlated with those

of the first half) described a theistic or absolutist position, emphasizing the immanence of God as the source of truth and moral law. There would have been equal justification for naming Factor III the "Theist" or "Absolutist" scale, on the one hand, or the "Relativist" or "Experimentalist" scale on the other. Because the last seemed the most appropriate option in an educational setting, the scale was named the *Experimentalist* factor.

As mentioned above, the first three factors of the School Opinion Survey appeared to describe three major dimensions of educational philosophy: the humanist, realist, and experimentalist. The intercorrelation between these three factors was nil; i.e., they seemed to represent orthogonal dimensions. In contrast to the first three factors, the remaining seven exhibited a fair degree of intercorrelation, while preserving their distinction as separate factors. These factors all represented the fourth plane of the theoretical model, the plane of educational techniques, i.e., the specific methods advocated to realize educational objectives. It is of interest to note that none of the technique items loaded significantly into any philosophical scale during factor analysis. This suggested that the educational techniques which people advocate may have much less relationship to the philosophy, values, and objectives they espouse than most would like to believe.

The seven technique scales were named respectively, *Individual Attention, Group Activities, Professionalization, Non-Academic, Academic Discipline, Scientific Objectivity*, and *Strict Control*. The first of the technique scales, Factor IV, stressed (1) individual attention and counseling and (2) closer home-school and parent-

TABLE 1

School Opinion Survey—Scale Intercorrelation Matrix

2	3	4	5	6	7	8	9	10	Factor	Scale Name
-.06	-.07	.33	-.07	.26	.18	.24	.19	.25	I	Humanist
	.04	-.24	.30	-.18	-.02	-.22	.13	-.14	II	Realist
		.06	.09	.15	.14	-.06	.11	-.18	III	Experimentalist
			.09	.52	.49	.49	.39	.47	IV	Individual Attention
				.08	.33	-.24	.26	-.22	V	Group Activities
					.33	.40	.43	.36	VI	Professionalization
						.02	.38	.03	VII	Non-Academic
							.20	.52	VIII	Academic Discipline
								.31	IX	Scientific Objectivity
									X	Strict Control

teacher relationships. Advocated were individual counseling on personal problems, individual attention to pupils and encouragement of creativity, parent orientation, parent-teacher conferences, home visits by teachers, school social workers, and better training of counselors. Common to all these characteristics was the concept of increased attention given by school staff to individual students in their school, classroom, and home educational environments, i.e., the "pupil personnel orientation." This scale correlated with the Humanist scale (.33) and negatively with the Realist scale (—.24). It also correlated well with each of the technique scales except Factor V. (Group Activities). It was decided to designate Factor IV the *Individual Attention* factor.

The fifth factor contained items emphasizing competitive and team sports, co-educational physical education, outdoor play and field trips, extracurricular activities and group projects, grading on the curve, PTA activities, and local school board autonomy. The outdoor, athletic, non-academic and small group emphasis were apparent. Any individual focus was conspicuously absent. Of all the technique factors, this correlated best with the Realist scale (.30). It also correlated with Non-Academic (.33) and Scientific Objectivity (.26) and negatively with Academic Discipline (—.24) and Strict Control (—.22), to be described below. Factor V was named the *Group Activities* factor.

Factor VI items advocated larger school districts and the aspects of staff professionalization usually associated with them: higher salaries for teachers and administrators, clerical help for teachers, educational research, school psychologists, and individual attention programs to assist the emotionally disturbed pupil and to encourage the gifted. It is not surprising to learn that Factor VI correlated with the Humanist scale and even slightly with the Experimentalist scale. Among the technique scales, it correlated well with Individual Attention, Academic Discipline and Scientific Objectivity, yet to be described. It correlated least with Group Activities and negatively with the Realist scale. Factor VI was designated the *Professionalization* scale.

Factor VII was composed of items stressing nonacademic and practical curriculum elements such as student government, shop and craft classes, art and music classes, consumer education, and other techniques aimed at preventing dropouts and retaining the

interest of poorly motivated or non-academically oriented students, e.g., better lighting and library, methods courses for teachers, and having incidental expenses of education assumed by the school. The essence of this factor seemed to be the special provisions in curriculum, school plant, services, and class presentation necessary to interest and retain non-academically oriented students. This factor correlated with both the Humanist and Experimentalist scales and also with Individual Attention, Group Activities, Professionalization, and Scientific Objectivity. It was uncorrelated with the Realist scale, Academic Discipline and Strict Control. Factor VII was named the *Nonacademic* scale.

Factor VIII items advocated increased time spent in study: more hours in the school day, weeks in school year, longer class periods and less free time. Also stressed were "solid" highly verbal college-preparatory type subjects involving concentrated study: foreign language, social studies, writing of themes and emphasis on great literature. This factor correlated positively with the Humanist scale and negatively with the Realist scale. Among the technique factors it correlated best with Strict Control, then with Individual Attention, Professionalization, and (negatively) with Group Activities. It was uncorrelated with Nonacademic. It seemed most appropriate to term this the *Academic Discipline* factor.

Factor IX was composed of items concerned with relatively impersonal scientific teaching methods: use of teaching machines, team teaching, stress on mathematics and use of personality, IQ, standardized and objective tests. Also included were state regulation of education, free medical care for students, and child-study training. This factor exhibited low positive correlations with all three philosophy scales and among the technique scales correlated in descending order with Professionalization, Individual Attention, Nonacademic, Strict Control, Group Activities, and Academic Discipline. The core concept of this factor was somewhat more difficult to determine than was that of the others. However, since the predominant element was the use of impersonal scientific educational methods, it has been designated the *Scientific Objectivity* scale.

The last factor was concerned with the theme of strict discipline and moral training: strict enforcement of school rules, strictness of discipline, spanking of misbehaving pupils, stringent laws against truancy, teaching of morals, and self discipline in school. Corporal

punishment was advocated and the belief expressed that punishment usually produces the desired results. This factor correlated moderately with the Humanist and negatively with the Realist and Experimentalist factors. It also correlated in descending order with Academic Discipline, Individual Attention, Professionalization, and Scientific Objectivity. There was no relationship with Nonacademic and moderate negative correlation with Group Activities. It seemed most appropriate to name this the *Strict Control* factor.

It is evident that Individual Attention is the most pervasive of the technique factors. It stems from a Humanist (and anti-Realist) philosophical position. This approach to education seems to be expressed in various ways: through increased Professionalization of the staff, special provision for Nonacademically oriented students, Strict Control of the fractious, and even to some degree through the application of Scientific Objectivity.

Over against Individual Attention (and practically uncorrelated with it) is Group Activities. Here the emphasis is on the social context of individuals rather than on any one individual himself; hence the correlation with the Realist and, even slightly, with the Experimentalist philosophical positions. The impersonal contextual emphasis of Group Activities is reflected in its correlated factors: Nonacademic and Scientific Objectivity. As expected, it correlates negatively with Academic Discipline and Strict Control. One is tempted to term this factor the "Playboy Orientation," as one of its major themes is to make school a sort of game.

Thus, there seem to be two major technique "families," distinguished primarily by whether the focus is on the individual student and his personal development or on the group context, group interaction and play. These orientations seem to be related, although imperfectly, to the humanist and realist philosophical positions respectively. A second but less prominent division signified by zero intercorrelation appears to lie between the *Nonacademic* factor on the one hand, and the factors of *Academic Discipline* and *Strict Control* on the other. The positive correlation of the former and negative correlations of the latter pair with the Group Activities and Experimentalist factors hint of the participation of the age-old relativist-absolutist philosophical controversy in this dimension.

Implications

Besides meeting the experimental necessity which gave it birth, the School Opinion Survey offers a means of measuring and describing diagnostically attitudinal differences among educational role groups. It is hypothesized that such data can help unearth the causes of such problems as relatively poor general academic performance, low morale, or poor human relations within a school system and/or poor rapport between a school system and its community. When a marked difference is found between the values or preferred techniques to which a school system, on the one hand, and the majority of parents in its community, on the other, subscribe, certain signs of stress may be anticipated. When a new principal takes over for another whose values and preferred techniques were far more congruent than are his with those of his faculty and the parents of that attendance area, difficulty may be just off-stage. It is conceivable that the availability of such measures will make possible not only the ferreting out of some causes of already developed difficulties but also the prediction and prevention of unnecessary values-techniques conflicts before they become entrenched. Data are at present being collected which will, it is hoped, shed light on the validity of these hypotheses. These data are expected to become available for publication during the coming year.

REFERENCES

Brackenbury, R. L. *Getting Down to Cases.* New York: G. P. Putnam's Sons, 1959.

Shaw, M. C. and Tuel, J. K. *A Proposed Model and Research Design for Pupil Personnel Services in the Public Schools.* Los Angeles: University of California Press, 1964. (a)

Shaw, M. C. and Tuel, J. K. Group Counseling with Parents. *California Journal of Educational Research,* 1964, 15, 232–249. (b)

Tuel, J. K. Exploring the Educational Environment. Address delivered to 91st Annual Forum of the National Conference on Social Welfare, 1964. (a)

Tuel, J. K. *The Educational Environment: A Useful Research Construct.* Guidance Research Project, University of California, 1964. (b)

Tuel, J. K. Major Philosophies in American Higher Education. *Improving College and University Teaching,* 1966, 14, 166–171.

Tuel, J. K. and Shaw, M. C. *School Opinion Survey*. Los Angeles: University of California, 1964.
Tuel, J. K. and Wursten, R. Dimensions of the Educational Environment. *California Journal of Educational Research*, 1965, 16, 175–188. (a)
Tuel, J. K. and Wursten, R. The Influence of Intra-personal Variables on Academic Achievement. *California Journal of Educational Research*, 1965, 16, 58–64. (b)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 965-972.

NEED-PRESS AND EXPECTATION-PRESS INDICES AS PREDICTORS OF COLLEGE ACHIEVEMENT[1]

CARL G. LAUTERBACH AND DAVID P. VIELHABER
Medical Research Project, West Point, N. Y.

WHEN a college applicant chooses a particular educational institution it may be assumed he does so largely because he expects it will most adequately satisfy his personal needs. If he is accepted he may or may not find the directive influences or "press" of the college in keeping with his needs or his expectations of the institution. It has been hypothesized that the extent of agreement (congruence) between one's internal forces (needs) and the external environmental forces (press) he encounters is positively related to his adaptation to that environment (Stern, 1960; Pace, 1961). This may be referred to as the need-press hypothesis. It may be, however, that it is not necessarily so much the congruence of his needs and the press as it is the congruence of what he expects (expectations) and the press he subsequently encounters that more strongly influences his adaptation. This may be referred to as an expectation-press hypothesis.

These hypotheses involve abstract concepts, i.e., press, needs, and expectations, and the defining and measurement of their referents pose certain difficulties; Getzels and Guba (1957), for example, have effectively stressed the need for establishing distinctions, classifying such terms, and stating their interrelationships with behavior. Of the three terms, the press of collegiate environments has perhaps been most successfully defined and measured, by such instruments as the College Characteristics Index (CCI) (Pace

[1] Impetus for this study stemmed directly from the original ideas and unfinished work of the late John P. Devlin; the authors are deeply indebted for his pioneering efforts. The study was supported by the U. S. Army Medical R & D Command and the United States Military Academy.

and Stern, 1958). However, research suggests that press alone is insufficient to predict students' achievements (Stricker, 1965), so that it may be more promising to utilize concepts such as need-press and expectation-press hypotheses to develop indices which include comparable representation of both internal forces (needs and expectations) and the press of the environment.

The purpose of this study was to make a comparative examination of the need-press and expectation-press hypotheses by empirically investigating need-press and expectation-press indices as predictors of adaptation of West Point cadets. These indices were derived by comparing profiles of the College Characteristics Index obtained from cadets under differing conditions.

Although it may be difficult to agree upon a sole criterion of a student's success in college, certainly in any collegiate environment academic achievement is a principal aspect of success. In addition, recognizing that a student's nonacademic performance is also an integral part of his total college development and achievement, West Point has developed a system whereby cadets are regularly rated by their peers and superiors regarding their leadership potential. In this study these Aptitude for Service Ratings, which have been found to be reliable measures and the best single predictor of later success as an officer (Crockett and Bowen, 1961), were considered in addition to grades as criteria of college achievement.

Method

Test Measures

The CCI, administered under special conditions, was used to obtain measures of new cadets' needs and expectations. It was also used to gauge the press of the institution which it traditionally measures. The same instrument was used for all of these purposes in order to diminish the problem of finding parallel scales from different instruments for comparisons of needs, expectations, and press.

Subjects and Conditions

383 cadets, the *N* or need group, were administered the CCI upon entrance to the Academy under special instructions to describe

West Point only as they preferred it to be (preference condition). Since they presumably had no first-hand experience with the institution, their individual CCI profiles were considered expressions of their desires or needs specifically related to their choice of collegiate environment. 387 other cadets, the *E* or expectations group, were also administered the CCI, but under the usual standard instructions to describe their institution as they saw it. Since this second group also had virtually no first-hand experience with West Point, their individual CCI profiles were considered expressions of their expectations. A measure of the press of the institution was derived from the mean CCI profile of 646 cadets from an earlier Plebe (freshman) class. They had completed the CCI under standard instructions midway through their Plebe year after they had experience with the environment. They are refererd to as the *P* or press group.

Need-Press Indices

For each of the 383 *S*s in the *N* group, two indices of need-press congruence were calculated. First, for a measure of the similarity of the *shape* of his need profile with the press of the environment, his 30 CCI scale scores were correlated with the mean scale scores of the *P* or press group. This correlation coefficient is referred to as his *r N-P* score. As an estimate of the *distance* between an *S*'s needs and the press of the environment, a *D* statistic (Cronbach and Gleser, 1953) was calculated between his 30 scale scores and the mean scale scores of the *P* group. This is referred to as his *D N-P* score.

Expectation-Press Indices

Similarly, for the 387 *S*s in the *E* group, two indices of expectation-press congruence were calculated. His *r E-P* score, an estimate of the similarity in *shape* of his expectations profile to the press of the environment, was the correlation coefficient between his 30 CCI scale scores and the 30 scale means of the *P* group. His *D E-P* score, an estimate of the *distance* between his expectations profile and the press, was the *D* statistic between his 30 scale scores and the scale means of the *P* group.

Criteria

The following indicators of USMA adaptation were obtained for all subjects in the *N* and *E* groups:

Academic Performance Criteria

Grade Point Average. An *S*'s average for all "academic" subjects, excluding physical education and strictly military subjects. This was obtained for his Plebe (freshman) year (GPA-1) and for his Second Class (junior) year (GPA-2).

Tactics Average. Grades for military subjects. They were designated TA-1 for Plebe year and TA-2 for Second Class year.

Nonacademic Performance Criteria

Aptitude for Service Ratings. Estimates of an *S*'s leadership potential based on ratings by his peers and superiors. They were obtained for the Plebe Year (ASR-1) and for the Second Class year (ASR-2).

Physical Education Average. This measure was also obtained for *S*'s Plebe year (PE-1) and the Second Class year (PE-2).

Results and Discussion

Neither of the need-press (*r N-P* and *D N-P*) nor the expectation-press (*r E-P* and *D N-P*) indices were correlated significantly with the nonacademic criteria (Physical Education average and Aptitude for Service Ratings) for either the Plebe or Second Class years. Therefore, as tested here, neither the need-press nor expectation-press hypotheses were useful in the prediction of non-classroom forms of cadet achievement.

However, positive support for the expectation-press hypothesis was gained from the correlations of the *r E-P* and *D E-P* indices with the four academic criteria (GPA-1, GPA-2, TA-1 and TA-2); all eight of these correlations were positive, and six were statistically significant. These results are shown in the first two columns of Table 1. None of them, however, was high; the coefficients ranged from .24 between *r E-P* and TA-1 to .11 between *r E-P* and GPA-2. The *D E-P* index appeared to have the most consistent validity of the two expectation-press indices, in that it correlated reliably with each of the four academic criteria, whereas the *r E-P*

index related significantly only with the two Tactics averages, TA-1 and TA-2. These results suggest that closeness in the *distance* between expectations and press profiles has more significance for predicting the academic criteria than does the similarity of their *shapes* alone. To summarize, the closer the expectations profile was to the press profile at West Point, the better an *S*'s subsequent academic achievement tended to be. Stated otherwise, the greater his accuracy or insight into the Plebe year press, the slightly greater the likelihood of his excelling in academic performance.

Of the corresponding eight correlations of the need-press indices (r *N-P* and *D N-P*) with the academic criteria, five were significant ($p < .05$). These results are reported in the last two columns of Table 1. However, the need-press indices correlated with each of the academic criteria in the *opposite* direction from that predicted from the need-press hypothesis! That is, all eight correlations were negative in sign, ranging from −.21 between the r *N-P* index and GPA-2, to −.06 between r *N-P* and TA-1. Thus, the less congruent an *S*'s CCI profile of preferences or needs was with the press, the better his academic achievement tended to be. Stated more directly, new cadets who preferred the West Point educational climate most differently from how experienced cadets reported it tended to be successful academically. Incongruity in the *shape* of the need and press profiles appeared to be about as predictive of grades as the *distance* between them, since both of the need-press indices correlated at the same significance level ($p < .01$) with GPA-1 and GPA-2.

TABLE 1

Correlations of Expectation-Press and Need-Press Indices with Freshman and Junior Year Academic Grades

Criteria	Index			
	r *E-P*	*D E-P*	r *N-P*	*D N-P*
	(N = 272)		(N = 263)	
GPA				
Freshman	.12	.19**	−.17**	−.15*
Junior	.11	.15*	−.21**	−.20**
Tactics Grades				
Freshman	.24**	.22**	−.09	−.06
Junior	.17*	.15*	−.13*	−.09

* $p < .05$.
** $p < .01$.

Since the need-press and expectation-press indices were surprisingly different in their relationships with academic criteria, an effort was made to understand better their meaning by determining their correlations with eight well-explored West Point selector variables: high school rank; an index of high school extracurricular activities; and SAT-Verbal, SAT-Math, Mathematic achievement, English composition, and physical aptitude test scores. The final selector variable used was the CEER score (College Entrance Examination (high school) Rank, which is a multiple regression composite of the five selector variables that best predicts West Point academic achievement. These results are shown in Table 2.

The correlations of the expectation-press indices with the eight selector variables were all positive, but only their correlations with SAT-Verbal (r's of .23 and .26) and English composition (r's of .21 and .23) were significant ($p < .01$). In contrast, both need-press indices tended to be negatively associated with SAT-Verbal, SAT-Math, and CEER scores, and were virtually unrelated to the other five selector variables. Only three significant correlations among the sixteen between need-press indices and selector variables were found: —.14 between r N-P and SAT-Math; —.15 between D N-P and SAT-Verbal; and —.15 also between D N-P and SAT-Math scores.

Since each of the expectation-press and need-press indices correlated in the same direction, and to about the same extent, with certain of the selector variables as they had with the academic

TABLE 2

Correlations of Expectation-Press and Need-Press Indices with Standard Selector Variables

Selector Variable	Index			
	r E-P (N = 272)	D E-P	r N-P (N = 263)	D N-P
High School Rank (Av.)	.10	.05	00	.01
Extra-Curricular	.04	.07	.03	.05
SAT-V	.26**	.23**	−.11	−.15*
SAT-M	.03	10	−.14*	−.15*
CEER	.04	.09	−.10	−.13
Mathematics	.07	.06	.06	.07
Eng. Comp.	.21**	.23**	.09	.00
Physical Apt. Test	.01	.05	.03	.03

* $p < .05$.
** $p < .01$.

criteria, it seemed likely that they were measuring certain cognitive factors in common with the selector variables. Such interpretation of the nature of the indices was supported by obtaining multiple correlations predicting academic grades using the CEER composite selector score in combination with the expectation-press and need-press indices. The results indicated that the correlations between CEER and Grade Point Average, which ranged from .42 to .63, could not be raised appreciably by combining any of the expectation-press or need-press indices with CEER. However, CEER alone correlated only .30 with TA-1 and .25 with TA-2, but when the *r E-P* index was combined with CEER, the resulting multiple correlation reached .38 for TA-1 and .30 for TA-2. Thus, it appears that the *r E-P* index may tend to contribute some unique variance to the specific prediction of Tactics (military subjects) grades not accounted for by the presently used selector variables. However, in general it appears that the expectation-press and need-press measures do not aid appreciably in the prediction of grades from presently available selector variables.

It may be concluded that the ability of a new cadet to anticipate accurately the press at West Point appears to be a function of cognitive factors. His initial knowledge of the educational environment at West Point can thus be viewed as a part of general verbal knowledge and ability which are also reflected in his SAT-Verbal and English Composition test scores.

However, the tendency of new cadets to express preferences which are incongruous with the actual press of the environment reported by experienced cadets was also related to cognitive factors as measured by selector variables. How cognitive factors operate in cadets' expression of their preferences relative to the West Point environment is not clear. It may be that intellectually outstanding cadets who score highest on particular selector variables (SAT-Verbal and SAT-Math) simply happen to have need-states which are most different from the press, but this need-press disparity is not relevant to their performance in the manner specified by the need-press hypothesis. Indeed, the empirical findings of this study indicate that close congruence between needs and press is associated with lesser academic achievement.

Perhaps a value of this study lies in its helping to denote some of the behavioral correlates of student expectations and needs or

preferences, required in order to comprehend the role of such concepts in affecting behavior within an institution, as urged by Getzels and Guba (1957). The results may also be suggestive of possible value in increasing efforts to provide early realistic orientation to new students, at least to West Point cadets, concerning the institutional realities which they will subsequently experience.

Summary

Half of an entering Cadet Class described West Point on the College Characteristics Index (CCI) as they saw it (Expectations), and half were instructed to describe it as they preferred it (Needs). The individual CCI profiles in each group were compared with the mean CCI profile of 646 experienced cadets (Press), resulting in individual need-press and expectation-press congruence indices. Correlations of these indices with academic grades lend support to the expectation-press hypothesis, but for the need-press hypothesis the results were opposite from those predicted. Relationships of the congruence indices with usual college selector variables suggest that they may be principally cognitive measures.

REFERENCES

Crockett, E. P. and Bowen, T. W. *The Aptitude for the Service System.* West Point, New York: Office of Military Psychology and Leadership, 1961 (Limited distribution).

Cronbach, L. J. and Gleser, G. C. Assessing Similarity between Profiles. *Psychological Bulletin*, 1953, 50, 456–473.

Getzels, J. W. and Guba, E. G. Social Behavior and the Administrative Process. *School Review*, 1957, 65, 423–441.

Pace, C. R. The Validity of the CCI as a Measure of College Atmosphere. Paper read at American Psychological Association Convention Symposium, New York City, September, 1961.

Pace, C. R. and Stern, G. G. An Approach to the Measurement of Psychological Characteristics of College Environments. *Journal of Educational Psychology*, 1958, 49, 269–277.

Stern, G. C. Congruence and Dissonance in the Ecology of College Students. *Student Medicine*, 1960, 8, 304–339.

Stricker, G. Intellectual and Nonintellective Correlates of Grade-point Average. Proceedings of the 73rd Annual Convention of the American Psychological Association. Washington, D. C.: American Psychological Association, 1965.

Educational and Psychological Measurement
1966, 26, 973–976.

A COMPARATIVE STUDY OF PERCEPTIONS OF A UNIVERSITY ENVIRONMENT BETWEEN HONOR AND NONHONOR FRESHMEN GROUPS

SHELDON R. BAKER
Wisconsin State University, River Falls

Present studies of college and university environmental characteristics indicate that major plans of study, level of training sought (Thistlewaite, 1960), and changes in level of aspirations (Thistlewaite, 1966) are significantly related to environmental press. Level of grade expectation and decisiveness in vocational decision making are also found to be significantly related to the perception of press (Baker, 1965).

In the above studies as in those of Astin and Holland (1961), Astin (1962), Stern (1963a,b), and Pace (1963), college environmental characteristics are conceptualized as global variables capable of shaping the productivity of academic institutions and as capable of affecting the decisions of students in their competition for limited academic rewards. This study accepts the paradigm which defines environmental press as a global factor, but only as it bears upon indigenous productivity, such as the election of an honors program, and the maintenance of academic status.

The election of an honors program by students generally indicates a strong need to learn as well as a need for intellectual recognition. Do students who elect an honors program perceive the press of their university significantly different as compared to those who do not select such a program? This study will attempt to consider the foregoing question by comparing the perceptions of environmental press between honors and nonhonors students.

With this purpose in view, the following null hypothesis is formulated for research:

There are no differences in the perception of environmental press between freshman honors program participants and freshman nonhonors participants.

Procedures

Fifty-seven second quarter freshmen honors program participants were administered the Stern's College Characteristics Index. Within one week another group of thirty-three freshmen nonhonors students volunteered to take the Index. The second group were part of a randomly selected number of one hundred students who were invited by letter to participate in the study. The second group of students had not elected an honors program nor were they asked by the administration to elect the program.

Thirteen factor scores were calculated for each student. The mean scores for each group were converted to standard scores with a mean of 0 and a standard deviation of 2. Appropriate t tests of significance were used to evaluate the null hypothesis, and the .05 level of significance was accepted as the minimum for the rejection of the null hypothesis.

TABLE 1

Standard Score Difference in Press Perception between Honors and Nonhonors Students, $N = 90$, $M = 0$, $SD = 2$

Factors	Nonhonors ($N = 33$)	Honors ($N = 57$)	Difference
Intellectual Climate Factors			
Work-Play	.5	−2.0	2.5**
Non-Vocational Climate	−1.3	−1.1	−.2
Aspirational Level	−2.1	−1.0	1.0*
Intellectual Climate	−1.8	−2.0	.2
Student Dignity	−2.5	−1.6	.9*
Academic Climate	−2.0	−2.0	0.0
Non-Intellectual Climate Factors			
Academic Achievement	−2.3	−1.7	−.6
Self Expression	−2.6	−1.2	−1.4**
Group Life	−.2	.8	−1.0*
Academic Organization	−.3	−.7	.4
Social Form	.7	.6	.1
Play-Work	1.1	2.0	−.9*
Vocational Climate	1.4	1.1	.3

* .01 level.
** .001 level.

Results

The major findings are summarized in Table 1. As can be seen from the data, the null hypothesis is rejected for differences between the following factors:

Work-Play	.001 level
Aspiration Level	.01 level
Self Expression	.001 level
Group Life	.01 level
Play-Work	.01 level
Student Dignity	.01 level

Honor students tend to perceive greater strength on Aspiration Level, Student Dignity, Self Expression, Group Life, and Play-Work as compared to nonhonor students.

The data also indicate that the learning environments within the same institution differ for honors as compared to nonhonors students.

Implications

The findings do not support the model of press as a global factor but suggest that an environment may contain multiple sets of press which become manifest upon the demands of salient needs of students in achieved status positions. Perhaps the variables which are purported to be measured by the College Characteristics Index are subject to the same types of influence as are social judgments and other types of social perceptions. The study recommends further elaboration with regard to the construct validity of the College Characteristics Index.

REFERENCES

Astin, A. An Empirical Characterization of Higher Educational Institutions. *Journal of Educational Psychology,* 1962, 53, 224–285.

Astin, A. and Holland, J. The Environmental Assessment Technique: A Way to Measure College Environments. *Journal of Educational Psychology,* 1961, 52, 308–316.

Baker, S. R. Vocational Indecisiveness and Decisiveness and Level of Grade Expectation as Related to Perception of University Environmental Press. *Perceptual and Motor Skills,* 1965, 21, 305–306.

Pace, C. *Cues, College and University Environmental Services.* Educational Testing Service, New Jersey, 1963.

Stern, G. Characteristics of the Intellectual Climate in College Environments. *Harvard Educational Review*, 1963, 33, 5–41. (a)

Stern, G. Scoring Instructions and College Norms. *Activities Index and College Characteristics Index*, Psychological Research Center, Syracuse, New York, 1963. (b)

Thistlewaite, D. College Press and Student Achievement. *Journal of Educational Psychology*, 1960, 51, 222–234.

Thistlewaite, D. and Wheeler, H. Effects of Teacher and Peer Subculture upon Student Aspiration. *Journal of Educational Psychology*, 1966, 57, 35–47.

Educational and Psychological Measurement
1966, 26, 977–983.

THE VALIDITY OF A COMPREHENSIVE COLLEGE SOPHOMORE TEST BATTERY FOR USE IN SELECTION, PLACEMENT, AND ADVISEMENT

THOMAS M. GOOLSBY, JR.
The Florida State University

It has become increasingly important in the past decade to provide for comprehensive testing at the college sophomore level on several counts. First, it is going to be almost imperative that selection occur for students entering the upper division. Secondly, it is very important to advise and guide students into appropriate curriculums of study in the upper division. Thirdly, there is the matter of shaping the curriculum of the junior college and the lower division of other institutions of higher learning.

At the present time there is no single test, battery of tests, or total testing program to provide adequate information for intelligent decision-making to meet any of the needs alluded to above.

It has been the experience of the admissions personnel and academic advisers that there is a dearth of adequate information for their needs. In most cases an estimate of success in school is based on grades or grade point averages, which are subsequently used as criteria. From grades in high school or in the lower division of the University, predictions of grades in future academic work have been made. From the scatter-plots of these relationships, expectancy tables have been developed for use in selection, placement, and advisement. Such expectancy tables have been based on single variable relationships rarely exceeding .60 and multiple-correlation coefficients rarely exceeding .75. It has not been infrequent that researchers have called for a thorough examination of the validity of grades or the grade point average. Intuitively, most researchers recognize that grades have poor measurement characteristics and

that any statistical manipulations based on them lead to many kinds of erroneous conclusions. According to Yonge (1965):

> Many investigators have been moving away from the old model of predicting intellective criteria (e.g., grades) from intellective predictors (e.g., ability tests) . . .
> Although grades are still a popular criterion, many investigators have displayed an interest in non-intellective predictors or correlates of grades . . .
> Goodstein and Heilbrun found that personality factors contribute most to the prediction of the academic achievement of the average college male . . .

Very little or no research effort has been made to develop a substitute measure for grades or a measure nearly so good as any of the predictors frequently used.

On the national scene, reputable measurement specialists have been concerned with sophomore testing and have been baffled by the complexity of the problems involved in making provisions for comprehensive measures. Among the experimental instruments is the Comprehensive College Tests (CCT) published by Educational Testing Service.

The Comprehensive College Tests were developed to provide a flexible program of college achievement testing to be used for a variety of purposes. Among these purposes are: evaluation of independent study for college credit, college equivalency, transition to upper-division studies, curriculum evaluation, and institutional self-study. The program of Comprehensive College Tests consists of the General Examinations and the Subject Examinations.

The General Examinations provide a comprehensive measure of undergraduate achievement in the five basic areas of liberal arts education: English Composition, Humanities, Mathematics, Natural Sciences, and Social Sciences-History.

An experiment was designed to attempt to determine to what extent the General Examinations of CCT could be used for (a) counseling students into the most appropriate upper division fields, and (b) screening students who might have academic difficulty in the upper division at a large southeastern university.

Procedures

The CCT was administered to currently enrolled second-semester sophomores in March, 1964. Florida Twelfth Grade Achievement

Tests (FTGAT) scores, School and College Ability Tests (SCAT) scores, and Sophomore Cumulative Grade Point Averages (Soph CGPA) were also collected in March, 1964. FTGAT and SCAT were administered in October, 1961. The Soph CGPA scores were collected by subject-matter area and total. Junior Year Grade Point Averages (Junior GPA) were collected a year later.

Intercorrelations among all measures were obtained. These intercorrelations were obtained for two purposes: to determine the measurement characteristics of CCT and to determine the utility of CCT as a predictor of certain criteria, including Grade Point Average (GPA).

Results

Table 1 indicates the correlations between CCT subtests. These relationships closely approximate those reported in other studies involving the use of CCT. The correlations between the subtests of Natural Science (.68), Mathematics (.64), and Humanities (.65) are of a magnitude indicative of some degree of independence. Without very much question, the subtests in these three areas are justifiable. The correlation between the subtests of the Social Sciences and History Test (.84) is considerably high. One may question the inclusion of these separate subtests in the CCT Battery. Subtests with such a degree of relationship may be justified on the basis of a desire to emphasize a variety of important objectives in the curriculum and to facilitate follow-up activities after the tests are taken. For purposes of selection and placement, however, the use of both of these tests is questionable.

TABLE 1

Correlations between CCT Subtests ($N = 700$)

CCT Subject Tests	Correlation
1. Natural Sciences Physical Sciences Biological Sciences	.68
2. Mathematics Basic Skills Content and Concepts	.64
3. Humanities Fine Arts Literature	.65
4. Social Sciences and History Social Sciences History	.84

From the intercorrelations presented in Table 2, one can immediately determine that there is a fair degree of independence among all tests of the CCT. Although these interrelationships were quite low, they were all significantly different from zero. As expected, the greatest degree of independence was between Mathematics and Humanities (.15), and the greatest amount of overlap was in English and Humanities (.56). The range of coefficients, however, was not very great (.15 to .56). In general, the degree of independence indicated here is somewhat unusual for achievement tests and highly desirable when the test battery is considered as a collection of entities.

In general, Table 3 indicates something of the predictive power of CCT. The highest relationship between Soph CGPA and any given subtest of CCT was .45. Even though this degree of relationship is significant, it has relatively little practical value for predictive purposes. When Junior Year GPA was used as a criterion, the highest relationship with any given test of CCT was .25. For practical uses, the tests of the CCT do not serve as a good predictor of Junior GPA or Sophomore CGPA.

TABLE 2

CCT Intercorrelations (N = 700)

CCT Subject Tests	English	Natural Sciences	Mathematics	Humanities	Social Sciences
English		.38	.32	.56	.31
Natural Sciences			.53	.39	.49
Mathematics				.15	.31
Humanities					.44

TABLE 3

Correlations of Certain Predictors of GPA (N = 700)

Predictor Variables	Soph CGPA	Junior GPA
SCAT Total	.40	.18
12th Grade Test	.37	.15
CCT English	.45	.25
CCT Natural Sciences	.30	.15
CCT Mathematics	.30	.11
CCT Humanities	.43	.23
CCT Social Sciences	.33	.13
Soph CGPA vs Junior GPA	.58	

Included in Table 3 is the relationship between Soph CGPA and Junior Year GPA (.58). The reduced magnitude of the relationship was somewhat due to the small variance of scores for the Sophomore CGPA.

As indicated in Table 4, the relationship between CCT and FTGAT was .75. One would consider such a relationship to be of some practical value. The relationship is interesting, since, for all intents and purposes, both test batteries were designed to measure instructional outcomes. This degree of relationship may be accounted for by considering the rigorous structuring of the tests according to specifications and the standard conditions under which they were administered. With a general improvement in testing and grading practices for courses taught in the University, one may hopefully expect a relationship between any of the instruments in Table 4 to correlate with GPA at least as high as the correlations among the tests; i.e., at least .70. When SCAT was used as a predictor of FTGAT, Soph CGPA, or CCT, the relationships were respectively .73, .50, and .70. SCAT does not predict Soph CGPA for any practical use. It does, however, predict FTGAT and CCT with a fair degree of efficiency. Here again, one may surmise that rigor in constructing and administering tests is a probable contributor to the higher relationships.

When certain tests of CCT were combined in a multiple prediction procedure as indicated in Table 5, the predictive efficiency improves very little.

Tables 6 and 7 present an analysis of CCT when subject area is of major concern. The means, standard deviations, and reliabilities of the CCT tests for the experimental group and the national norming group are presented in Table 6. In general, the means for the experimental group were slightly higher than the means for the

TABLE 4

Correlations of Certain Tests and of CCT with Soph CGPA ($N = 407$)

SCAT Total vs FTGAT	.73
SCAT vs CCT	.70
CCT vs FTGAT	.75
CCT vs Soph CGPA	.50

TABLE 5

CCT Multiple Correlations with GPA ($N = 700$)

Combinations of Predictor Variables	Soph. CGPA	GPA Jr. Year
CCT English + CCT Humanities	.45	.40
CCT Mathematics + CCT Humanities	.53	.24
CCT Natural Sciences + CCT Mathematics	.35	.21
CCT English + CCT Math + CCT Social Sciences	.45	
CCT Nat. Sci. + CCT Math. + CCT Social Sciences	.33	

national norm group. The standard deviations were in general, much lower for the experimental group than for the national norm group. Reliability coefficients reported in the manual range from .91 to .95 for the tests of the CCT.

As indicated in Table 7, the coefficients of correlation between CCT and certain subject areas are quite low for any practical predictive use.

A superficial inspection of a scatter plot, not presented here, will indicate something of the inefficiency of CCT to predict Soph CGPA. Approximately an equal number of students who achieved below a GPA of 2.01 were equally distributed at each score from 324-606 on the CCT English. When considering students who made a GPA of 2.01, there is still one chance in three that the true

TABLE 6

Means, Standard Deviations, and Reliabilities for CCT

	CCT Subject Tests				
	Eng.	N.S.	Ms.	Hs.	SS & Hy
Mean					
Exp.	512.4	503.3	475.7	530.0	515.1
Nat.	498	498	498	499	498
SD					
Exp.	76.7	76.0	54.7	74.8	75.5
Nat.	99.0	99.0	99.0	99.0	99.0
r_{11}	.92	.91	.95	.91	.92

Exp. — Experimental Group in Present Study
Nat. — National Sample
Eng. — English
N.S. — Natural Sciences
Ms. — Mathematics
Hs. — Humanities
SS. — Social Sciences
Hy. — History
r_{11} — Reliability

TABLE 7

Correlations between Parts of the CCT and GPA by Subject Area (N = 407)

CCT Subject Test Predictors	GPA Criteria					
	Eh. GPA	N.S. GPA	Ms. GPA	Hs. GPA	SS & Hy. GPA	Soph CGPA
CCT Eh.	.56	.28	.19	.20	.34	.39
CCT N.S.	.18	.36	.05	.12	.12	.26
CCT Ms.	.08	.33	.27	.04	.23	.27
CCT Hs.	.47	.22	.05	.16	.53	.28
CCT SS & Hy.	.10	.22	.05	.16	.53	.28

GPA — Grade Point Average
Eh. — English
N.S. — Natural Sciences
Ms. — Mathematics
Hs. — Humanities
SS — Social Sciences
Hy. — History

Soph CGPA is greater than 2.42 or less than 1.60. This magnitude of error in prediction leaves much to be desired in making any sort of practical decision about selection, placement, and advisement.

Summary and Conclusions

A comprehensive sophomore college testing program is becoming imperative for selection, placement, and advisement purposes.

The CCT has satisfactory measurement characteristics on most counts and excellent ones on a few. The evidence presented in this paper does not support the use of the CCT for selection, placement, and advisement at the sophomore level when it is considered alone and especially when GPA is a criterion. Furthermore, there is a need to question the desirability of the use of GPA as a criterion.

There is evidence in this paper to support a need for a very substantial emphasis on adequate measurement and grading practices within the University.

A rigorous determination and definition of curriculum and the construction of criterion measures (cognitive and affective) in a continuing research program is necessary for higher education to meet its responsibilities in selection, placement, and advisement.

REFERENCE

Yonge, George D. Students. *Review of Educational Research,* 1965, 35, 253–263.

Educational and Psychological Measurement
1966, 26, 985–995.

MULTIPLE DISCRIMINANT PREDICTION OF COLLEGE CAREER CHOICE

RALPH B. VACCHIANO[1]
Fairleigh Dickinson University

AND

ROBERT J. ADRIAN
Queen's College, City University of New York

Research studies have revealed that there exist distinct personality patterns which characterize specific professional groups (Izard, 1960; Roe, 1946a, 1946b, 1951a, 1951b, 1953; Schaffer, 1953; Siegelman and Peck, 1960) as well as specific academic groups preparing for various professional fields (Merwin and DiVesta, 1959; Garrison and Scott, 1961; Stern, 1962; Stern, Stein and Bloom, 1956; Zuckerman, 1958). Prior investigations have emphasized the differentiation and description of these professional and academic groups. Few studies though, have attempted to predict group membership, either professional or academic, based on the personality dimensions they have described.

The present study departed from the more frequently adopted approach of univariate analysis and, utilizing mulitple discriminant analysis, explored the feasibility of predicting students' academic choice based on personality need constructs as measured by Stern's Activity Index.

Method

Sample

Three male groups representing three academic areas of study, business, chemistry, and mathematics, and two female groups, edu-

[1] The authors wish to express their appreciation to the members of the Applied Mathematics Division, Esso Research and Engineering Company, and in particular to J. Beckwith, M. Efroymson, R. Hardy, and R. Heitler.

cation and nursing, served as criterion groups. These groups were comprised of a total of 245 students, 50 in each group, with the exception of chemistry which contained 45 students. Two additional male groups, 32 business and 36 mathematics students, and two additional female groups, 32 education and 38 nursing students, were utilized as cross-validation samples to test the classification efficiency of the discriminant functions derived from the criterion groups. Although the groups were pre-selected according to their area of study, students were assigned randomly to the criterion and cross-validation samples. The students shared several characteristics in common. They were all Caucasian, native-born American, approximately 21 years of age and, with the exception of the nursing students, all were matriculated, day students attending a large metropolitan university. Nursing students were drawn from a metropolitan hospital-affiliated nursing school. At the time of testing, all students were seniors, and all had indicated that they intended to pursue the professional career for which they were then preparing. Students who had changed their major academic area of study more than once, who had changed majors because of academic difficulty, who had indicated dissatisfaction with their choice of major or who were not maintaining satisfactory academic standing were eliminated from the sample.

Predictor Variables

Subjects were administered the Activities Index (AI) in group sessions lasting approximately one hour. The AI consists of 300 items describing commonplace daily activities and feelings which are keyed to 30 of the Murray needs. There are 10 items for each of the 30 measured needs to which the subject is required to respond for each item with a "like" or "dislike" choice (Stern, 1958). All 30 variables were used in the analysis.

Three multiple discriminant analyses[2] were computed, one for the three male groups of business, chemistry, and mathematics students, one for the same business and mathematics groups (omitting chemistry students), and one for the two female groups of education and

---

[2] The analysis was carried out by means of DISCRIM, a program developed by Cooley and Lohnes (1962) and adapted for use with the IBM 7094 by Jones (1964).

nursing students. Classification was based on the derived discriminant functions.[3]

Results

The means, univariate F tests, and scaled vectors are summarized in Table 1 for the male business, chemistry, and mathematics groups. The descriminating power of the predictor test battery was determined by computation of Wilks' lambda, which is a function of the roots of $W^{-1}A$, where W is the pooled within-groups deviation scores cross products matrix, and A is the among-groups matrix of cross products of deviations of group from grand means weighted by group sizes (Cooley and Lohnes, 1962; Rao, 1952). The effectiveness of this discrimination was significant ($F = 2.25$, df $= 60/225$, $p < .001$). Chi square tests (Bartlett, 1941) were computed for each of the two derived discriminant functions (by removal of successive latent roots) to determine the significance of discrimination along each dimension (Jones, 1964). The first discriminant function was found to be significant ($\chi^2 = 120$, $p < .001$), but the significance of the second vector failed to reach the necessary level ($p = .11$). The first vector accounted for 72 per cent of the predictable group variation.

The relative contribution of the predictor variables to each of the discriminant functions may be seen from an examination of the scaled weights (Table 1). These weights were derived by multiplication of the normalized latent vectors by the square root of the corresponding diagonal element of the W matrix. The scaled weights for the first discriminant vector indicate that the predictors providing the greatest contribution were determined by the business group's high scores for need affiliation, exhibition, and humanism; the chemistry students' high need scores for practicality and science; and the mathematics students' high need scores for succorance.

Since the second vector proved to be nonsignificant, difficulty was encountered not only in interpretation of factors, but also in later classifications. Another discriminant analysis was then computed for the same criterion group of business and mathematics students, omitting the chemistry group. For this analysis, the ef-

[3] Classification was accomplished by means of CLASSIF (Cooley and Lohnes, 1962), adapted for use with the IBM 7094 by Jones (1964).

fectiveness of the discriminating power of the test, Wilks' lambda, was significant ($F = 2.56$, $df = 29/70$, $p < .001$) as was the discriminating power of the one discriminant function[4] computed ($\chi^2 = 60.78$, $p < .001$). The function accounted for 100 per cent of the predictable group variation.

A review of the scaled weights summarized in Table 2 indicates that many of the predictor variables which contributed to the

TABLE 1

Means, Univariate F Tests, and Scaled Vectors for Male Business, Chemistry and Mathematics Groups

Needs	Means: Bus	Chem	Math	F*	p	Scaled Vectors: I	II
Abasement	3.72	3.82	3.98	.25	ns	.08	.09
Achievement	6.38	5.67	5.38	2.50	.08	.05	−.06
Adaptability	4.92	5.22	4.92	.33	ns	−.09	−.15
Affiliation	7.10	5.69	6.12	4.46	.01	.18	−.03
Aggression	5.22	4.64	4.06	3.06	.04	.00	−.18
Change	5.26	4.27	4.70	2.95	.05	−.06	.07
Conjunctivity	4.60	6.09	5.50	5.06	.007	−.06	−.01
Counteraction	6.50	6.33	6.20	.27	ns	.10	−.14
Deference	5.36	6.11	6.08	2.19	ns	−.14	−.04
Dominance	6.64	6.36	5.36	4.01	.01	−.09	−.25
Ego Achievement	5.80	4.96	3.96	5.36	.005	.02	.04
Emotionality	3.18	3.58	3.88	1.95	ns	−.21	−.06
Energy	6.90	6.33	6.84	1.53	ns	−.02	.27
Exhibitionism	4.78	3.78	3.28	4.51	.01	.18	.02
Fantasized Achievement	5.18	4.78	3.88	3.99	.02	.01	−.38
Harm Avoidance	3.26	4.82	4.52	5.62	.004	−.09	.03
Humanism	5.12	5.09	4.20	1.92	ns	.40	−.24
Impulsiveness	5.74	4.13	4.90	8.79	.0003	.16	.20
Narcissism	4.68	4.42	4.78	.29	ns	.06	.09
Nurturance	5.34	5.38	5.46	.03	ns	.05	−.08
Objectivity	8.60	8.24	8.56	1.11	ns	.17	.10
Order	4.08	5.78	4.74	5.03	.007	.06	−.36
Play	6.04	4.96	5.54	2.45	.08	−.05	−.29
Practicality	5.68	6.24	5.72	.94	ns	.15	−.01
Reflectiveness	6.06	6.89	6.46	1.72	ns	−.02	.10
Science	3.58	7.31	5.90	25.17	.0001	−.71	−.19
Sensuality	4.26	4.18	4.54	.55	ns	.02	.24
Sex	3.96	4.20	4.84	1.71	ns	−.13	.10
Succorance	5.60	6.20	6.80	4.69	.01	−.23	.22
Understanding	6.54	6.93	7.30	1.53	ns	.00	.35

* df = 2/142.
Bus = Business
Chem = Chemistry
Math = Mathematics

[4] The number of discriminant functions computed is g-1, where g = groups.

TABLE 2

Means, Univariate F Tests, and Scaled Vectors for Male Business and Mathematics Groups

Needs	Means		F*	p	Scaled Vectors
	Bus	Math			I
Abasement	3.72	3.98	.48	ns	−.01
Achievement	6.38	5.38	5.62	.01	.14
Affiliation	7.10	6.12	5.23	.01	.21
Aggression	5.22	4.06	5.56	.01	.13
Change	5.26	4.70	2.00	ns	−.07
Conjunctivity	4.60	5.50	3.47	.01	−.03
Counteraction	6.50	6.20	.51	ns	.07
Deference	5.36	6.08	2.74	.01	−.13
Dominance	6.64	5.36	7.76	.006	.09
Ego Achievement	5.80	3.96	10.64	.001	−.02
Emotionality	3.18	3.88	3.86	.04	−.10
Energy	6.90	6.84	.03	ns	−.09
Exhibitionism	4.78	3.28	9.03	.003	−.01
Fantasized Achievement	5.18	3.88	7.39	.007	.26
Harm Avoidance	3.26	4.52	6.75	.01	−.09
Humanism	5.12	4.20	3.13	.07	.50
Impulsiveness	5.74	4.90	4.60	.03	.11
Narcissism	4.68	4.78	.04	ns	−.15
Nurturance	5.34	5.46	.06	ns	.17
Objectivity	8.60	8.56	.04	ns	.19
Order	4.08	4.74	1.44	ns	.29
Play	6.04	5.54	1.19	ns	.02
Practicality	5.68	5.72	.01	ns	.07
Reflectiveness	6.06	6.46	.79	ns	−.21
Science	3.58	5.90	17.34	.0001	−.35
Sensuality	4.26	4.54	.63	ns	−.10
Sex	3.96	4.84	3.23	.07	−.01
Succorance	5.60	6.80	9.19	.003	−.25
Understanding	6.54	7.30	3.11	.07	−.31

Note.—Because of programing difficulties, only 29 need variables were utilized in this analysis.
* $df = 1/98$.
Bus = Business
Math = Mathematics

differentiation between the three male groups were also responsible for the differentiation between the business and mathematics groups for this analysis. The mean need scores for the business and mathematics students remain the same in this analysis. When consideration was given to the means of all three male groups (Table 1), the mean scores of the chemistry students tended to fall between those of the business and mathematics students. When the mean scores for the chemistry students are no longer considered (Table 2), the contrast between the business and mathematics students

becomes more apparent. The scales furnishing the greatest relative contribution were determined by the business group's high scores for needs affiliation, aggression, dominance, exhibition, fantasized achievement, humanism, and play, and the mathematics group's high scores for needs order, science, succorance, and understanding. The ambivalence of the motivations of the business students was suggested by the findings that these students appear to seek out close, friendly relationships, while at the same time expressing feelings of aggression and ascendency over others. Their interest in people would seem to be a reflection of their own need for self-gratification, since their social participation expresses their needs for self-display and attention seeking. Achievement motivation tends to be on a fantasized level revolving around daydreams of success in achieving extraordinary public recognition. In contrast, mathematics students receive satisfaction in imposing order and in manipulating their environment symbolically. These students would seem to have a greater need for detached intellectualization through utilizing problem solving or theorizing as an end in themselves. In interpersonal situations, though, the mathematics students' affectional needs are satisfied by dependency on others.

Table 3 summarizes the results obtained with the two female groups. Wilks' lambda proved to be significant ($F = 2.07$, df $= 30/69$, $p < .006$) as did the discriminating power of the discriminant function ($\chi^2 = 53$, $p < .005$). The function accounted for 100 per cent of the predictable group variation. The scaled weights indicate that the following variables yielded the greatest relative contributions: dominance, ego achievement, energy, harm avoidance, humanism, narcissism, objectivity, and sex. Female education students displayed greater needs for dominance or ascendency, a striving for power, fearfulness, interest in the humanities and social sciences, narcissistic trends, and objectivity which was in contrast to the nursing students who displayed greater needs for activity and heterosexual interests. The fact that two need variables, narcissism and sex, with insignificant univariate F tests, contributed to the discriminant function, may be explained by the value of employing multivariate techniques. The contribution of these variables was relatively high in terms of the factor loadings for the total profile (see Cooley and Lohnes, 1962, pp. 119–121, for a discussion of this phenomenon).

TABLE 3

Means, Univariate F Tests, and Scaled Vectors for Female Education and Nursing Students

Needs	Means Educ	Means Nurs	F*	p	Scaled Vectors I
Abasement	4.08	4.02	.03	ns	.07
Achievement	4.56	4.62	.02	ns	.25
Adaptability	5.56	5.66	.05	ns	.07
Affiliation	7.86	7.76	.10	ns	−.06
Aggression	2.80	2.82	.00	ns	.08
Change	5.74	5.92	.22	ns	.01
Conjunctivity	5.20	5.08	.09	ns	.06
Counteraction	5.66	6.34	3.14	.07	−.21
Deference	6.64	6.54	.06	ns	−.03
Dominance	5.26	3.78	11.61	.002	.39
Ego Achievement	4.86	3.86	4.12	.04	.19
Emotionality	5.40	5.52	.10	ns	.18
Energy	6.04	6.78	4.59	.03	−.29
Exhibitionism	3.30	2.90	.83	ns	−.02
Fantasized Achievement	2.90	2.84	.02	ns	−.21
Harm Avoidance	5.58	4.78	3.80	.05	.20
Humanism	6.54	5.62	3.11	.07	.15
Impulsiveness	5.62	6.00	1.10	ns	−.30
Narcissism	5.72	5.08	2.19	ns	.25
Nurturance	7.38	7.40	.00	ns	.00
Objectivity	8.58	8.26	3.49	.06	.32
Order	5.08	5.22	.07	ns	−.19
Play	6.48	6.18	.70	ns	.09
Practicality	5.10	5.44	.64	ns	.02
Reflectiveness	6.08	6.04	.01	ns	−.14
Science	3.50	3.10	.52	ns	.06
Sensuality	5.08	5.12	.01	ns	−.03
Sex	6.28	6.46	.20	ns	−.33
Succorance	7.14	6.80	1.30	ns	.09
Understanding	5.36	5.54	.16	ns	−.11

* *df* = 1/98.
Educ = Education
Nurs = Nursing

Classification

Table 4 contains three contingency tables for predicted group membership (criterion groups), a 3 × 3 table for male business, chemistry, and mathematics subjects, a 2 × 2 table for male business and mathematics students, and a 2 × 2 table for the female education and nursing groups. Group membership was determined by the maximum likelihood classification of the individual's discriminant score. This method necessitates the computation of sep-

arate chi squares (one for each group used in the multivariate discriminant analysis) for each subject through using the centroids and dispersions for each group in combination with the individual's scores. Each chi square indicates the probability of that subject's membership in each group employed in the discriminant analysis (see Cooley and Lohnes, 1962, for a discussion of this procedure).

Since the three chi squares (Table 4) for both the male and female groups were significant ($p < .001$), the hypothesis of the independence of actual and predicted group membership was rejected. For the classification of the male business, chemistry, and mathematics groups, 74% of the business, 69% of the chemistry, and 56% of the mathematics students were correctly classified. Classification based on the second discriminant analysis, computed for only the business and mathematics students, resulted in the correct classification of 90% of the business and 88% of the mathematics students. For the female groups, 82% of the education and 84% of the nursing students were correctly classified.

An indication of the stability of the discriminant functions was obtained by applying the obtained weights to the cross-validation groups of male and female students. Three cross-validation tables are summarized in Table 5. Since the chi squares for the business and mathematics and education and nursing students were significant ($p = .05$ and $p = .02$), the independence of classification hypothesis was again rejected. This hypothesis could not be rejected for the stability of the weights derived from the combined business, chemistry, and mathematics groups ($p = .08$).

The correct classification of 65% of the education and 58% of the nursing students suggests that the female weights are relatively stable. Unfortunately, the same conclusion could not be drawn for the cross-validation male subjects when the discriminant analysis was applied to business, chemistry, and mathematics students. Only 54% of the business and 33% of the mathematics students were correctly classified. This is accounted for, in part, by the decreased probability of "hits" because of the lack of a cross-validation chemistry group. If mathematics and chemistry students are considered similar ("science" as opposed to business), then they may be combined into one group. Following this procedure, 88% of the "science" students were correctly classified in the criterion group and 72% of the "science" students were correctly classified in the cross-validation sample.

TABLE 4

Classification of Criterion Groups

		Predicted Group Membership								
		Male Groups						Female Groups		
		Bus	Chem	Math		Bus	Math		Educ	Nurs
Actual Group Membership	Bus	37	9	4	Bus	45	5	Educ	41	9
	Chem	5	31	9	Math	6	44	Nurs	8	42
	Math	6	16	28						
		$\chi^2 = 77.24$				$\chi^2 = 63.68$			$\chi^2 = 43.56$	
		$p = <.001$				$p = <.001$			$p = <.001$	
		$df = 4$				$df = 1$			$df = 1$	

Bus = Business
Chem = Chemistry
Math = Mathematics
Educ = Education
Nurs = Nursing

TABLE 5

Classification of Cross-validation Groups

		Predicted Group Membership								
		Male Groups						Female Groups		
		Bus	Chem	Math		Bus	Math		Educ	Nurs
Actual Group Membership	Bus	18	7	8	Bus	21	12	Educ	21	11
	Math	10	14	12	Math	13	23	Nurs	16	22
		$\chi^2 = 5.26$				$\chi^2 = 5.20$			$\chi^2 = 3.82$	
		$p = .08$				$p = .02$			$p = .05$	
		$df = 2$				$df = 1$			$df = 1$	

Bus = Business
Chem = Chemistry
Math = Mathematics
Educ = Education
Nurs = Nursing

Application of the discriminant weights obtained from the multiple discriminant analysis of business and mathematics students only, indicated that the derived weights were relatively stable. Sixty-four per cent of the business and 64% of the mathematics students were correctly classified.

Discussion

Predictions of group membership remained above the base rate for both the male business and mathematics cross-validation groups and female education and nursing cross-validation groups. The findings would suggest that prediction of academic major based on the personality variables measured by the Activities Index is feasible. This observation does not exclude the fact that

other cognitive factors are not important components in determining vocational goals. The inability of the discriminant weights to differentiate successfully between male chemistry and mathematics students would indicate that such variables are in operation.

Although the high mean need scores for the groups in the present study paralleled those needs which Stern (1962) reported as significantly separating similar academic groupings, these variables did not contribute to the classification of students in the present study. This was true for the education and nursing groups, particularly for needs narcissism and sex. This finding is due to the analysis utilized, namely, the multivariate approach. In such a procedure it is possible for two groups to have insignificant univariate F tests on a variable; yet the factor loadings may be sufficient to aid in classification. Application of the mulivariate technique, in contradistinction to the more frequently employed univariate tests, also takes into consideration the combined effects of a number of variables. Thus classification efficiency is maximized at no expense to the description of group dynamics.

The present investigation also serves to demonstrate that prediction of college career choice based on need patterns may be quite fruitful. Future research may adopt a similar approach in combining other variables with measures of personality in order to maximize classification efficiency and to aid the college student in his choice of college career plans.

REFERENCES

Bartlett, M. S. The Statistical Significance of Canonical Correlations. *Biometrika*, 1941, 32, 29–38.

Cooley, W. and Lohnes, P. *Multivariate Procedures for the Behavioral Sciences*. New York: Wiley, 1962.

Garrison, K. and Scott, M. A Comparison of the Personal Needs of College Students Preparing to Teach in Different Teaching Areas. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1961, 21, 955–964.

Izard, C. Personality Characteristics of Engineers as Measured by the Edwards Personal Preference Schedule. *Journal of Applied Psychology*, 1960, 44, 332–335.

Jones, K. J. *The Multivariate Statistical Analyzer*. Cambridge, Massachusetts: Harvard Coop., 1964.

Merwin, J. and DiVesta, F. A Study of Need Theory and Career Choice. *Journal of Counseling Psychology*, 1959, 6, 302–308.

Murray, H. A. and Collaborators. *Explorations in Personality*. New York: Oxford, 1938.

Rao, C. R. *Advanced Statistical Methods in Biometric Research*. New York: Wiley, 1952.

Roe, A. The Personality of Artists. Educational and Psychological Measurement, 1946, 6, 401–408. (a)

Roe, A. A Rorschach Study of a Group of Scientists and Technicians. *Journal of Consulting Psychology*, 1946, 10, 317–327. (b)

Roe, A. A Psychological Study of Eminent Physical Scientists. *Genetic Psychological Monographs*, 1951, 43, 121–239. (a)

Roe, A. A Psychological Study of Eminent Biologists. *Psychological Monographs*, 1951, 65, No. 14 (Whole No. 331). (b)

Roe, A. A Psychological Study of Eminent Psychologists and Anthropologists. *Psychological Monographs*, 1953, 67, No. 2 (Whole No. 352).

Schaffer, R. E. Job Satisfaction as Related to Need Satisfaction in Work. *Psychological Monographs*, 1953, 67, No. 14 (Whole No. 364).

Siegelman, M. and Peck, R. Personality Patterns Related to Occupational Roles. *Genetic Psychological Monographs*, 1960, 61, 291–349.

Stern, G. G., Stein, M. and Bloom, B. *Methods in Personality Assessment*. Glencoe, Illinois: The Free Press, 1956.

Stern, G. G. *Preliminary Manual Activities Index—College Characteristics Index*. Syracuse, New York: Syracuse University, 1958.

Stern, G. The Measurement of Psychological Characteristics of Students and Learning Environments. In S. Messick and J. Ross, (Eds.), *Measurement in Personality and Cognition*. New York: Wiley, 1962, pp. 27–68.

Zuckerman, M. The Validity of the Edwards Personal Preference Schedule in the Measurement of Dependency-Rebelliousness. *Journal of Clinical Psychology*, 1958, 14, 379–382.

Educational and Psychological Measurement
1966, 26, 997–1005.

THE RELIABILITY AND VALIDITY OF A NEW MEASURE OF LEVEL OF OCCUPATIONAL ASPIRATION[1]

BERT W. WESTBROOK
North Carolina State University

To meet the need for a practical, reliable, and valid measure of level of occupational aspiration (LOA), Haller (1957) has developed The Occupational Aspiration Scale (OAS). Although LOA has been recognized as a factor of some significance in the social mobility of youth (Empey, 1956; Haller, 1959; Haller and Miller, 1963; Holloway and Berreman, 1959; Rosen, 1959; Sewell, Haller, and Straus, 1957), research has been restricted considerably by the lack of an instrument for measuring it. Evidence collected from several samples of male high school students suggests that the OAS is a valid measure of LOA (Miller and Haller, 1964). The purpose of this paper is to examine the characteristics of the OAS resulting from administrations of the instrument to eleventh grade boys and girls.

Level of Occupational Aspiration

Miller and Haller (1964) define LOA as the area of the occupational prestige hierarchy which an individual views as a goal. The range of an individual's LOA is determined by what he views as realistically probable versus idealistically desirable for him, and by the goals which he has for the near versus the distant future (Lewin, Dembo, Festinger, and Sears, 1944). Both the realistic or idealistic goals and the time boundaries must be specified in order for a subject to respond adequately to an LOA instrument.

[1] This paper is partly based upon a dissertation submitted by the author in partial fulfillment of the requirements for the doctor's degree at Florida State University, Tallahassee.

The Occupational Aspiration Scale

Haller's OAS has been described completely elsewhere (Haller and Miller, 1963). It is a multiple-choice instrument designed primarily for use among male high school students. Previously this instrument had not been employed with a sample of high school girls. The OAS takes into consideration idealistic and realistic levels at particular future time periods, and it combines these variables in such a way as to yield a single score for each person. Questions are asked at each of two career points: short-range (S) and long-range (L) for the realistic level (R) and for the idealistic level (I). These four variables produce four combinations: RS (realistic-short), IS (idealistic-short), RL (realistic-long), IL (idealistic-long). The combinations are incorporated into four wordings for questions. The question wordings are as follows: IS—of the jobs listed in this question, which ONE would you choose if you were FREE to CHOOSE ANY of them you wished when your SCHOOLING IS OVER? (questions 2 and 4); IL—Of the jobs listed in this question, which ONE would you choose to have when you are 30 YEARS OLD, if you were FREE TO HAVE ANY of them you wished? (questions 6 and 8); RS—Of the jobs listed in this question, which is the BEST ONE you are REALLY SURE YOU CAN GET when your SCHOOLING IS OVER? (questions 1 and 3); RL—Of the jobs listed in this question, which is the BEST ONE you are REALLY SURE YOU CAN HAVE by the time you are 30 YEARS OLD? (questions 5 and 7). Since each of the four questions is presented twice, there is a total of eight OAS items.

The response alternatives for each OAS item consist of 10 occupational titles drawn from among the 90 occupations ranked by the National Opinion Research Center's study of the prestige of occupations (National Opinion Research Center, 1947). The ten least appropriate occupations were dropped, leaving a total of 80 occupations, ten for each question. The 80 occupations were ranked from first to eightieth. The occupational title with the highest prestige rank was placed in the first question, the title with the second highest rank was placed in the second question, and so on down to the occupational title with lowest prestige rank which was placed in the last question.

Each question contains a set of occupational alternatives which

span the full range of occupational prestige. It is scored from zero to nine. Each occupation is presented only once among the eight items. If the respondent chooses the highest prestige occupation in a given question item, he receives a score of 9: if he chooses the lowest prestige occupation, he receives a score of 0. The sum of the scores on each of the eight questions may range from zero to seventy-two. The total score represents the individual's level of occupational aspiration.

The OAS takes between fifteen and twenty minutes to administer. It may be administered easily in a group testing situation. It is recommended that the meaning of the occupational titles should not be explained to the respondents. This is necessary in order to reduce error due to difference among administrators and to simulate the reality situation in which a student must choose among the alternatives he knows (Miller and Haller, 1964).

Sample and Data

The OAS was administerd by the investigator to a total sample of 154 eleventh grade students—77 boys and 87 girls during the spring of 1964. Additional data from this sample included scholastic ability results and a questionnaire regarding future plans of each student. The OAS was readministered after a two-week interval and again after a five-week interval to both girls and boys. The distributions of total OAS raw scores for the three administrations are approximately normal.

Table 1 shows the means, standard deviations, and ranges for both boys and girls on each administration of the OAS. Although there was no significant difference in mean OAS total scores be-

TABLE 1

Means, Standard Deviations, and Ranges of OAS Scores for Boys and Girls on the First, Second, and Third Administration

	Boys			Girls		
Administration	Mean	S.D.	Range	Mean	S.D.	Range
First	45.78	10.84	23–67	45.47	7.61	30–60
Second (Two-week Interval)	45.67	11.13	25–67	46.11	7.23	32–59
Third (Five-week Interval)	45.94	12.02	23–68	44.84	8.02	28–61

tween boys and girls for any of the administrations, the girls' OAS scores were less variable than the boys. Both the ranges and standard deviations of the girls were smaller than those of the boys for each administration of the OAS. This finding can be explained by the fact that the occupations at the extreme ends of the OAS hierarchy are those for men primarily. Therefore, girls were forced to choose occupations closer to the center of the occupational prestige hierarchy.

Reliability

The reliability of the OAS was determined by computing coefficients of internal consistency (Spearman-Brown) based on the correlation between parallel halves, and by examining the test-retest correlations after two-week and five-week intervals to establish coefficients of stability. The design of the OAS makes it possible to determine the correlation between parallel halves since each half includes one complete set of question types (RS, RL, IS, and IL). The coefficients of internal consistency are as follows: first administration, .85; second administration, .83; third administration, .86.

The test-retest correlations provided coefficients of stability between the same form of the test administered two weeks and five weeks apart. The coefficient of stability for the two-week interval was .84 for males and .88 for females. The coefficient of stability for the five-week interval was .82 for males and .78 for females. The standard error of measurement was computed for each administration of the OAS. These were found to be 6.11, 5.89, and 5.33 for the first, second, and third administrations respectively. The findings show that the OAS is a fairly reliable instrument. The halves of the instrument appear to be equivalent, and the stability is reasonably high over a five-week period.

Validity

The evidence supporting the validity of the OAS comes from the internal structure of the instrument and from the correlation of the OAS total score with other measures. Level of occupational aspiration theory contends that level of aspiration at the idealistic level is higher than level of aspiration at the realistic level, and that level of aspiration for long-range goals should be higher than

level of aspiration for short-range goals (Lewin, Dembo, Festinger, and Sears, 1944). In the OAS, realistic (R) questions are designed to tap a lower limit of the respondent's LOA and idealistic (I) questions are designed to tap an upper limit of the respondent's LOA. Since the occupational achievement level of an individual is usually expected to rise to some extent during the first decade or so of his career, we would predict that long-range (L) LOA should be, on the average, higher than short-range (S) LOA. Specific tests of these hypotheses are as follows:

1. The realistic LOA should be less than the idealistic LOA. The mean OAS-RS score should be less than the mean OAS-IS score and the mean OAS-RL score should be less than the mean OAS-IL score.

2. The short-range LOA should be lower than the long-range LOA. The mean OAS-RS score should be less than the mean OAS-RL score and the mean OAS-IS score should be less than the mean OAS-IL score.

The data from the administration of the OAS partially supported the hypotheses. The mean OAS-RS score (4.18) is less than the mean OAS-IS score (4.50) and the mean OAS-RL score (4.10) is less than the OAS-IL score (4.19). Since the mean OAS-RS score (4.18) is not less than the OAS-RL score (4.10), and since the mean OAS-IS score (4.50) is not less than the mean OAS-IL score (4.19), the second hypothesis was not supported. Significance tests were not carried out, since each subtest was judged to measure a different characteristic.

To determine the factorial structure of the OAS, the eight OAS item scores were intercorrelated by the product-moment method and the resulting correlations were factor analyzed by the principal-axes technique. Five factors accounted for 93 per cent of the variance; this factor matrix was rotated orthogonally by the Quartimax method and by the Varimax method. The Quartimax solution is shown in Table 2.

Inspection of Table 2 shows that all eight OAS questions have moderately high loadings on Factor I. This finding tends to support the contention that each question of the OAS contributes to the measurement of a general LOA variable, at least in terms of the factor model used. None of the loadings on Factors II and III are above .40. Factor I accounts for 73 per cent of the total variance

TABLE 2

Quartimax Rotation of Factor Matrix of Item Scores on the Occupational Aspiration Scale

Question–Item	Factors				
	I	II	III	IV	V
1. Realistic–Short Range	54	05	09	42	07
2. Idealistic–Short Range	59	−30	−34	−10	28
3. Realistic–Short Range	71	05	−05	26	11
4. Idealistic–Short Range	63	−11	−19	09	19
5. Realistic–Long Range	68	30	11	−13	−04
6. Idealistic–Long Range	64	02	24	−27	00
7. Realistic–Long Range	61	26	08	03	09
8. Idealistic–Long Range	66	−20	16	−04	−13
Cumulative Proportion of Total Variance	73	80	86	90	93

Note.—All decimal points omitted.

and might be referred to as general level of occupational aspiration.

Theoretically, the nature of level of aspiration leads one to expect that group factors would emerge from items in the OAS. The group factors might be called realistic-short, idealistic-short, realistic-long, and idealistic-long. In an attempt to obtain a factor solution which would account for the hypothesized group factors, the original factor matrix was rotated by the Varimax method. It can be seen in Table 3 that the Varimax solution accounts for the group factors to a greater extent than the Quartimax solution did.[3] The two highest loadings on each factor are found on those questions which presumably measure identical dimensions of aspiration level. The two questions with the highest loadings on Factor II are questions dealing with realistic-long range goals; these are questions five and seven. The two idealistic-long range questions, questions six and eight, are highest on Factor III. Factor IV loads highest on questions one and three; both questions pertain to realistic-short range occupational choices. Although Factor V does not contain an appreciable loading for any of the questions in the OAS, the questions pertaining to idealistic-short range goals were

[3] In a personal communication to the investigator concerning rotational methods of factor analysis, Dr. William B. Michael suggested that the Varimax method might be used to obtain a more satisfactory solution than the Quartimax method. This suggestion is greatly appreciated, since the Varimax rotation produced a factor matrix which more nearly supports the hypothesized structure of occupational aspiration level.

TABLE 3

Varimax Rotation of Factor Matrix of Item Scores on the Occupational Aspiration Scale

Question–Item	Factors I	II	III	IV	V
1. Realistic–Short Range	49	08	−01	43	05
2. Idealistic–Short Range	53	−18	−21	−06	36
3. Realistic–Short Range	61	04	00	38	14
4. Idealistic–Short Range	55	−10	−17	13	29
5. Realistic–Long Range	57	44	10	−05	−07
6. Idealistic–Long Range	50	12	33	−16	12
7. Realistic–Long Range	51	42	07	−11	15
8. Idealistic–Long Range	54	−17	28	08	−11
Cumulative Proportion of Total Variance	73	80	86	90	93

Note.—All decimal points omitted.

found to have the highest loadings. Many of the factor loadings do not reach the .40 level; nevertheless, the pattern of loadings is consistent with the hypothesized structure suggested by level of aspiration theory and research.

Relationship between OAS and Other Variables

Additional evidence of the validity of the OAS was collected by examining the correlation of the total score with another instrument presumed to be a measure of level of occupational aspiration. An open-ended measure of LOA, known as the North-Hatt technique, was administered to all pupils in the present study. Miller and Haller (1964) found this instrument to be correlated with level of educational and occupational attainment several years later (.46) and with number of years completed at college (.52). The respondent is asked to indicate the occupations he would enter if he were absolutely free to go into any kind of work he wanted, the occupations he has thought about going into, the occupation he plans to follow, and the type of work he would like to be doing by the time he is 30 years old. The average prestige scores of the responses to these questions have a meaning similar to that of the OAS. When the two sets of scores obtained in this study were correlated, it was found that the two instruments correlated to the extent of .68.

A number of additional variables were found to be correlated with OAS scores. The correlations were as follows: number of years

of college planned (.69); high school grade-point average (.57); scholastic ability-SCAT (.45); parental desire for respondent's educational achievement (.48); and socio-economic status of the respondent's family (.43).

Summary

The evidence collected in this study strongly suggests that the OAS is a reliable and valid measure of level of occupational aspiration. Test-retest reliability coefficients over a period of two weeks and five weeks were found to be satisfactory. Hypotheses about the elevation of mean OAS item scores were partially supported. Although boys and girls did not differ significantly on the OAS total scores, the girls' scores showed less variability.

The factorial structure of the OAS derived from a Varimax rotation of the factor matrix supports the existence of both a general LOA factor and a number of group factors suggested by level of aspiration theory and previous research. The OAS should continue to be a useful instrument for studying level of occupational aspiration among high school pupils.

REFERENCES

Empey, LaMar T. Social Class and Occupational Aspiration: A Comparison of Absolute and Relative Measurement. *American Sociological Review*, 1956, 21, 703–709.

Haller, A. O. *The Occupational Aspiration Scale*. East Lansing, Michigan: The Social Research Service, Michigan State University, 1957.

Haller, A. O. Planning to Farm: A Social Psychological Interpretation. *Social Forces*, 1959, 37, 263–268.

Haller, A. O. and Miller, I. W. *The Occupational Aspiration Scale: Theory, Structure, and Correlation*. East Lansing, Michigan: Michigan State University, Technical Bulletin 288, 1963.

Holloway, R. G. and Berreman, J. V. The Educational and Occupational Aspirations and Plans of Negro and White Male Elementary School Students. *Pacific Sociology Review*, 1959, 2, 56–60.

Lewin, K., Dembo, Tamara, Festinger, L. and Sears, Pauline S. Level of Aspiration in J. McV. Hunt (Ed.), *Personality and the Behavior Disorders*, Vol. I. New York: Ronald Press, 1944, 333–378.

Miller, I. W. and Haller, A. O. A Measure of Level of Occupational Aspiration. *Personnel and Guidance Journal*, 1964, 42, 448–455.

National Opinion Research Center. Jobs and Occupations: A Popular Evaluation. *Opinion News*, 1947, 9, 3–13.

Rosen, B. C. Race, Ethnicity, and Achievement. *American Sociological Review*, 1959, 24, 47–60.

Sewell, W. H. Haller, A. O. and Straus, M. A. Social Class and Level of Educational and Occupational Aspiration. *American Sociological Review*, 1957, 22, 67–73.

Westbrook, Bert W. *The Effect of Test Reporting on Self-Estimates of Scholastic Ability and on Level of Occupational Aspiration of Eleventh-Grade Boys*. Doctor's thesis. Tallahassee: Florida State University, 1964. 103 pp. Abstract, *Dissertation Abstracts*, 1965.

Educational and Psychological Measurement
1966, 26, 1007–1013.

FURTHER VALIDATION OF A SCALE TO MEASURE PHILOSOPHIC-MINDEDNESS

DONALD W. FELKER
Institute for Child Study
University of Maryland

This study is concerned with validation of the PM Scale which is designed to measure philosophic-mindedness (PM). It was originally presented by Felker and Smith (1966).

The PM Scale has 84 items, in three sections. Each section is composed of 28 items; 14 deal with educational material and 14 deal with general material. The items are all forced-choice with two alternatives.

Section one of the scale presents four paragraphs describing situations in which teachers might find themselves. After each paragraph the testee is presented with a series of paired statements. He is asked to choose the statement in each pair describing what he would be more likely to do if faced by the situation described in the paragraph.

Section two asks the testee to choose the statement in each pair which describes more accurately a person whose judgment he would respect.

Section three asks the testee to choose the statement in each pair with which he more closely agrees.

Analysis of philosophic-mindedness was begun by Smith (1956). A person is philosophic-minded whose thinking is characterized by three dimensions, i.e., comprehensiveness, penetration, and flexibility. The PM Scale was developed using flexibility as its criterion attribute. Philosophic-minded flexibility was postulated as having four characteristics. These characteristics are:

1. Being free from psychological rigidity.
2. Ability to evaluate ideas apart from their source.
3. Seeing issues as many-sided rather than two-sided, and the development of a relatively large number of alternate hypotheses, explanations, or viewpoints.
4. Maintaining a tolerance for tentativeness and suspended judgment and a willingness to take action in an ambiguous situation.

The PM Scale yields a total score and four subscores. The subscores are individual measures of these four characteristics. For the present study the total score appeared to be the most relevant, and hence the analysis of the subscores is not reported.

Review of Previous Study

The original study involving the scale dealt with the construction of items, the general characteristics of the scale, and hypotheses which were tested to begin construct validation of the scale.

When the test was administered to 358 subjects enrolled in graduate courses in education, 76 items were correlated at the .01 level of significance with the criterion scores, and two items were correlated at the .05 level of significance with the criterion scores. The remaining six items had positive correlations which were not significant.

The PM Scale exhibited a test-retest reliability coefficient of .80 with approximately three months between testings. The Spearman-Brown split half reliability coefficient was also .80.

The hypotheses which were tested in the previous study to begin construct validation of the scale and the results were:

1. If the scale measures PM, it will significantly separate students who are rated on flexibility by their instructors in philosophy of education courses. Those rated high on flexibility scored significantly higher on the PM Scale than those rated low. The differences were significant at the .001 level.

2. If the scale measures PM the scores will have a negative relationship to dogmatism and closed-mindedness. The correlation between the PM Scale and Rokeach's Dogmatism Scale was —.393.

The analysis of the findings of this study tended to support the scale as being a valid measure of PM.

Purpose

The purpose of the present study was to continue the investigation of the validity of this scale. Three hypotheses were presented and tested. These hypotheses and their relationship to the construct are:

Hypothesis One: If the test measures PM, there should be an increase in scores subsequent to taking courses in philosophy of education that have this attribute as their aim.

Although Felker and Smith (1966) presented this hypothesis, it was not tested. PM is an attribute that is changeable by learning experiences. There should be a general increase in scores associated with experiences designed to increase the attribute. The present study tested this hypothesis by using a group before and after it had taken an introductory course in philosophy of education.

Hypothesis Two: If the scale measures PM, there should be a significant positive correlation between the pretest PM score and success in the course as measured by course grades.

Since "Introductory Philosophy of Education" deals with questions which are judged to require the types of skills and intellectual dispositions which are characteristic of PM, it would follow that the person who possesses a greater measure of these characteristics should be better prepared for the course. Although these students should be better prepared, one would predict only a moderate correlation because of the changes that would take place over the course and because of the limitation posed by the reliability of the criterion.

Hypothesis Three: If the scale measures PM, there should be a significant positive correlation between PM scores on the posttest and success in the course as measured by course grades.

The grades in the course were computed from two types of assignments. Each student was given two objective tests, and two papers (one on a subject of the student's choice, the other on an assigned topic). Both papers were assigned with the purpose of displaying the characteristics of PM. The student was graded on the basis of the degree of PM displayed in attacking the problem of the paper.

The aim of this course was the cultivation of PM, and the grades in the course were meant to be a measure of the attainment of this

goal. In both the course grades and the PM Scale, one should have two experimentally independent measures of PM for which a positive correlation should be observed.

Procedures

Two classes at Indiana University in "Introductory Philosophy of Education" were each randomly divided into two groups. Two of these groups were given the PM Scale as a pretest. One of the groups was released from the class time when the PM Scale was administered. The fourth group wrote a brief essay on the subject "How will philosophy of education help me in my work?" while the pretest was being given to the rest of the class. All of the subjects were given the PM Scale at the end of the course.

These procedures were followed in order to test Hypothesis One. Two alternate explanations might be given for an increase in scores from the pretest to the posttest. These alternate explanations would be: (a) The increase is due to taking the test a second time, or (b) The pretest would make the student more aware of what was going to be presented in the course and hence would interact with the course to produce an increase in scores over the course.

A group of 103 students who were not enrolled in a philosophy of education course showed no increase in scores from a first to a second administration. This would make the first alternate explanation unacceptable.

In order to test the second explanation an analysis of variance was computed for the posttest scores of the four experimental groups. If gains in scores could be explained by the taking of the pretest and then the course, a significant difference between those who had taken the pretest and those who had not taken the pretest should have appeared on the posttest. The analysis of variance of the posttest scores showed no significant difference between the groups.

For the testing of the three validation hypotheses which were proposed in the study the subjects were divided into two groups; those who had taken the pretest and those who had not taken the pretest. There were 47 subjects in each group.

The purpose of the pre- and posttests of PM was not explained to the subjects until all had taken the posttest. The PM Scale was not scored until all course work was completed and the grades cal-

culated. The course tests and papers were graded on a nine-point scale, A+ to C−. The total point value of the two course tests and the two course papers was then used in computing the correlations in this study.

Since all members of the experimental group were under the same instructor in "Introductory Philosophy of Education," a second group under a different instructor was used for cross-validation purposes. There were 31 subjects in this group.

Results

Hypothesis One: If the test measures PM, there should be an increase in scores associated with taking courses in philosophy of education that have this attribute as their aim.

TABLE 1

Repeated Measures Analysis of Variance of PM Scores on Pretest and Posttest for Experimental Group and Cross-Validation Group

Group	Source	df	SS	MS	F	Sign. Level
Experimental	Between	1	1744.95	1744.95	49.08	.001
	Residual	46	1635.55	35.56		
Cross-Validation	Between	1	248.00	248.00	10.81	.01
	Residual	30	688.00	22.93		

An analysis of variance was computed for PM scores on the pretest and the posttest for the group enrolled in philosophy of education. This analysis is presented in Table 1. Table 1 also presents the analyses of variance for pretest and posttest PM scores for the cross-validation group. A repeated measures analysis was used for both of these analyses of variance. Both of these groups showed significant gains in PM scores subsequent to the course. These analyses would tend to support the original hypothesis.

Hypothesis Two: If the scale measures PM there should be a significant positive correlation between the pretest PM score and course grades. The PM Scale and the course grades for the pretest group had a Pearson correlation of .507. For the cross-validation group, the correlation of the same variables was .506. These correlations were significant beyond the .01 level and would support the hypothesis.

Hypothesis Three: If the scale measures PM there should be a

significant positive correlation between the posttest PM score and the course grades.

This hypothesis was tested for three groups: those who had taken a PM pretest, those who had not, and the cross-validation group. It was not known whether the taking of the pretest would be related to any significant differences in the correlations. The correlation between PM and grades in the group that had taken the pretest was .559. The correlation between PM and the grades in the group that had not taken the pretest was .767. These two correlations are not significantly different at the .05 level. Both of the correlations, however, are significant beyond the .01 level.

The correlation between PM and grades in the cross-validation group was .513 which is significant beyond the .01 level.

These correlations would support the hypothesis.

Discussion and Summary

Three hypotheses were proposed in the process of construct validation of the PM Scale. The hypotheses were tested using 94 graduate students enrolled in a beginning course in philosophy of education. This course, which had the cultivation of PM as its aim, would seem to be suited to testing the hypotheses.

The hypotheses were also tested on another class of 31 subjects for cross-validation purposes.

Each of the hypotheses received positive support from the analyses. Differences between pre- and post-administration of the PM Scale were significant at the .001 level. The PM scores, both the pretest and the posttest, were significantly correlated with course grades.

Although no measure of general intelligence was obtained on the subjects in this study, a previous study by Felker and Smith (1966) showed a correlation of .421 between a general measure of intelligence and the PM Scale. Although this correlation is significant, it accounts for only a small portion of the variance in the PM Scale. Because of this it would not appear that the results of our present hypotheses would be adequately explained on the basis of general intelligence.

The hypotheses tested in this study and the results observed would continue to support the PM Scale as a valid measure of the stated attribute of philosophic-mindedness.

REFERENCES

Felker, Donald W. and Smith, Philip G., *The Measurement of Philosophic-Mindedness on the Criterion of Flexibility,* Bulletin of the School of Education, Indiana University, Vol. 42, No. 1 (Jan.) 1966, 138 pp.

Smith, Philip G., *Philosophic-Mindedness in School Administration,* College of Education, The Ohio State University, Columbus, Ohio, 1956, 129 pp.

Educational and Psychological Measurement
1966, 26, 1015–1020.

THE RELATIONSHIP OF THE 1960 REVISED STANFORD-BINET INTELLIGENCE SCALE TO INTELLIGENCE AND ACHIEVEMENT TEST SCORES OVER A THREE-YEAR PERIOD

WILLIAM D. CHURCHILL and STUART E. SMITH
Alfred University

All reviews of the third revision of the Stanford-Binet Intelligence Scale (1960 S-B) essentially agree that the latest revision represents an advance in technical quality over the 1937 edition. However, it was noted by most of the reviewers that its predictive validity had not been ascertained at the time of publication. Balinsky (1960), pp. 155–156) noted that "the probability is high that the validity of the revision will be at least equal to if not greater than the 1937 version" because of the great amount of overlap between the two revisions. Fraser's review in Buros (1965, pp. 830–831) indicates that "there is no guarantee that the final test, taken as a whole, will give the same results as either the old Form L or Form M." Subsequent to 1960, only a few validity studies have been published. For a sample of elementary school children, Estes, et al. (1961) reported a correlation of .82 between the 1960 S-B and the 1937 S-B. In this same study a correlation of .74 between the 1960 S-B and the WISC (full scale) was reported. Tatham and Dole (1963) reported a correlation of .41 between the 1960 S-B and the California Test of Mental Maturity-SF for one group of elementary students; for a second group they reported a correlation of .56.

Purpose

The primary objective of this study was to determine the relationship between the 1960 S-B and the Lorge-Thorndike IQs ob-

tained three years later. A secondary objective was to determine the relationship between the 1960 S-B and achievement as measured by a standardized achievement test battery.

Method

At the time the criterion data were collected for this study, the subjects were members of the seventh grade in a rural central school of New York State. When these students were in the third grade, the 1960 S-B was administered to each member of the class. There were 82 members in the class, 45 boys and 37 girls. Because of attrition over a three-year period, the authors were able to compile complete data on 56 members of this group, 32 boys and 24 girls.

The data were obtained from the school's records of test information on file for each subject in the sample, as follows:

1. IQ, 1960 S-B, administered grade 3
2. Composite score, Iowa Test of Basic Skills, Form 2, 1955 edition, administered grade 3, fall semester
3. Composite score, Iowa Test of Basic Skills, Form 1, 1955 edition, administered grade 6, fall semester
4. IQ, Lorge-Thorndike Intelligence Tests, Verbal and Nonverbal batteries, Form A Level 4, administered grade 6, fall semester

The third grade 1960 S-B IQs were correlated with the sixth grade Lorge-Thorndike verbal IQ (L-T verbal) and the Lorge-Thorndike nonverbal IQ (L-T nonverbal). Correlations were obtained also between the 1960 S-B, L-T verbal, L-T nonverbal, and the Iowa Test of Basic Skills (ITBS) achievement scores (composite) obtained in grade three and grade six. The intercorrelations among the tests are shown in Table 1, together with the respective means and standard deviations. Tabulations of changes in scores between the 1960 S-B and the L-T verbal and L-T nonverbal were accomplished by inspection, with the results as indicated in Tables 2 and 3.

Results

The Mean 1960 S-B for this sample of 56 students was 110.25 compared with an L-T verbal mean of 110.92. The obtained standard deviation for the 1960 S-B is 15.8 which corresponds closely to the

standard deviation of 16.0 reported by Terman and Merrill (1960, p. 28). The L-T standard deviations are somewhat smaller (12.5 and 14.1) than the reported value of 16 (Lorge and Thorndike, 1954, p. 3).

TABLE 1

Means, Standard Deviations, and Intercorrelations for 1960 Stanford-Binet, Lorge-Thorndike, and Iowa Tests of Basic Skills

Test	$\bar{X}$	SD	S–B	L–T V	L–T NV	ITBS 3	ITBS 6
S–B	110.25	15.89	—	.79	.65	.59	.74
L–T (V)	110.92	12.52		—	.61	.62	.84
L–T (NV)	113.44	14.19			—	.47	.65
ITBS (3)	3.50	.77				—	.79
ITBS (6)	6.59	1.24					—

The primary objective of this study was to determine the predictive validity of the 1960 S-B over a three-year period using the L-T IQs as criteria. For the 56 students, the obtained correlation between the third grade 1960 S-B and the L-T verbal was .79, and the correlation between the third grade 1960 S-B and L-T nonverbal was .65.

Table 2 shows the amount of change in IQs on the L-T verbal compared with the 1960 S-B. Exactly one-half of the sample, 28 students (including one who registered no change), obtained IQs in the range ± 5 points of their third grade 1960 S-B IQs. Three students obtained L-T verbal IQs which were 16 to 25 points above. One student showed no change in IQ between the 1960 S-B and the L-T verbal. (This person's score has been omitted in Table 2.)

Table 3 shows the amount of change in the L-T nonverbal IQs compared with the third grade 1960 S-B IQs. Eighteen students (including two who showed no change) obtained L-T nonverbal

TABLE 2

Increase or Decrease in L–T Verbal IQs from 1960 S–B IQs for 55 Students over a Three-year Period

Score change	1–5	6–10	11–15	16–20	21–25	26–30	Total
Increase	12	8	5	1	2	0	28
Decrease	15	5	4	2	0	1	27
Total	27	13	9	3	2	1	55/55

IQs in the range of ± 5 points of their third grade S-B IQs. One student's IQ was 30 points higher than his third grade 1960 S-B IQ, while one student's L-T nonverbal IQ was 36 points lower than his third grade 1960 S-B IQ. A majority, 34 students, scored higher on the L-T nonverbal than they did on the 1960 S-B. This finding reflects the higher L-T nonverbal mean of 113.44 as compared with the 1960 S-B mean of 110.25. Two students showed no change between the 1960 S-B and the L-T nonverbal. (These two students have been omitted from Table 3.)

The correlation between the 1960 S-B and the third grade ITBS is .59, the correlation between the 1960 S-B and the sixth grade ITBS is higher, .74.

TABLE 3

Increase or Decrease in L–T Nonverbal IQs from 1960 S–B IQs for 55 Students over a Three-year Period

Score change	1–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	Total
Increase	9	7	9	6	2	1			34
Decrease	7	8	1	1	2			1	20
Total	16	15	10	7	4	1		1	54/54

The intercorrelations among the four criterion variables are presented in Table 1. For this sample, the intercorrelation between the two parts of the Lorge-Thorndike is .62. The highest obtained correlation in Table 1 (.84) is between the L-T verbal and the sixth grade ITBS, although it is only slightly higher than the next highest correlation of .79 (1960 S-B vs. L-T verbal). The third grade ITBS scores correlate .79 with the sixth grade ITBS scores. This value is identical with the obtained correlation of .79 between the 1960 S-B and the L-T verbal over the same period of time.

Discussion

The correlations obtained in this study between the 1960 S-B and the L-T verbal was .79. This correlation, based upon two different tests over a three-year interval, is approximately the same order of magnitude as for similar correlations reported for the 1937 revision of the Stanford-Binet (1937 S-B). For example, a correla-

tion of .82 has been reported (Krugman et al., 1951) between the 1937 S-B (form L) and the WISC (full scale). The correlation of the 1937 S-B (form L) with the WISC verbal scale was .74, and with the WISC performance scale the correlation was .64. It might be noted, however, that these correlations were based on scores obtained over a period of only a few months. It seems reasonable to assume that the correlations between the 1937 S-B and the WISC over a three-year period would be somewhat lower than those reported in the Krugman study.

The correlation between the 1960 S-B and the L-T nonverbal is .65, somewhat lower than the correlation of .79 between the 1960 S-B and the L-T verbal. This parallels the findings reported by Krugman et al., in which the WISC performance IQs correlated .64 with the 1937 S-B (form L).

The correlation of the 1960 S-B with the ITBS shows obtained correlations of .59 for grade three and .74 for grade six. It is unusual to find concurrent validity—third grade vs. third grade—to be lower than predictive validity—third grade vs. sixth grade.

Inasmuch as the four separate correlations involving the 1960 S-B and the four tests of "scholastic ability" correlate between .59 and .79, it appears that the 1960 S-B is a valid measure of "scholastic ability." Cronbach (1960, pp. 224-225) points out that "for most children, a group mental test leads to the same prediction that a comprehensive achievement test would." Referring to Table 1, it is noted that the correlation between the L-T verbal and the sixth grade ITBS is high (.84); the correlation of the L-T nonverbal and the sixth grade ITBS, while somewhat lower, is still quite high (.65). The correlation between the 1960 S-B and the sixth grade ITBS (.74) is of the same relatively high order as the correlation between the third grade ITBS and the sixth grade ITBS (.79). These results would seem to indicate that, for this sample, an individual mental test, a group mental test, and a comprehensive achievement battery are comparable in their predictive ability. These findings support Cronbach's generalization that group mental tests and comprehensive achievement tests lead to the same prediction.

Correlation coefficients are typically used to express or index the validity of intelligence tests. Another useful way to show the relationship between two tests is to show what range of IQs exist on a

second test for a given IQ interval on the first test. Although an expectancy table is not presented, the data in Table 2 may be helpful in this respect. One-half of the students obtained L-T verbal IQs which were not more than ± 5 points of their 1960 S-B IQs obtained three years earlier. Although the variability of the L-T verbal score at grade six is somewhat less than the variability of the 1960 S-B scores at grade three, the mean scores are practically identical. Based upon the data for this small sample, it appears that the two IQs can be used practically interchangeably.

Although the sample used in this study was small, the results presented here indicate that the latest revision of the Stanford-Binet is as useful a predictor of scholastic achievement as the 1937 revision. Further studies, involving larger and more diverse samples, are needed to establish the predictive validity of the 1960 S-B.

REFERENCES

Balinsky, B. Review of Stanford-Binet Intelligence Scale Manual for the Third Revision Form L-M. *Personnel and Guidance Journal*, 1960, 39, 155–156.

Cronbach L. J. *Essentials of Psychological Testing*. (2nd ed.). New York: Harper, 1960.

Estes, Betty W., Curtin, Mary E., DeBurger, R. A. and Denny, Charlotte. Relationships between the 1960 Stanford-Binet, 1937 Stanford-Binet, WISC, Raven, and Draw-A-Man. *Journal of Consulting Psychology*, 1961, 25, 388–391.

Fraser, Elizabeth D. Review in O.K. Buros (Ed.), *The Sixth Mental Measurements Yearbook*. Highland Park, New Jersey: The Gryphon Press, 1965. Pp. 830–831.

Krugman, Judith I., Justman, J., Wrightstone, J. W., and Krugman, M. Pupil Functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology*, 1951, 15, 475–483.

Lorge, I. and Thorndike, R. *General Manual the Lorge Thorndike Intelligence tests*. New York: Houghton Mifflin, 1954.

Tatham, C. B. and Dole, A. A. A Note on the Relationship of CTMM-SF to the Revised Binet, Form L-M. *Journal of Clinical Psychology*, 1963, 19, 302.

Terman, L. M. and Merrill, Maud A. *Stanford-Binet Intelligence Scale: Manual for the Third Revision Form L-M*. Boston: Houghton Mifflin, 1960.

Educational and Psychological Measurement
1966, 26, 1021–1029.

PREDICTION OF THE EMPLOYABILITY OF STUDENTS IN A SPECIAL EDUCATION WORK-TRAINING PROGRAM USING THE PORTEUS MAZE TEST AND A RATING SCALE OF PERSONAL EFFECTIVENESS[1]

SALVATORE GAMBARO and ROBERT E. SCHELL
Michigan State University

The present investigation is best described as an empirical prediction study. It was concerned with the use of the Porteus Maze test as a predictor of employability among educationally handicapped adolescents. The question prompting the study may be phrased in this form: Can a dependable cutting score be derived on the Porteus Maze test that will separate those individuals who will later be hired from those who will not?

The predictive relationship between rated personal characteristics and employability was also investigated. Comparison of this relationship with that of the Porteus test and employability made it possible to evaluate whether the better selection procedure would have been to use the Porteus test alone or in combination with another instrument.

Method

Subjects

The *Ss* were 71 adolescents—25 girls and 46 boys—between 16 and 19 years of age. This sample constituted all of those from a Lansing Special Education Department population of 84, for whom complete and usable test information was obtained.[2] All of the *Ss*

[1] Thanks are due Dr. Marvin Kaplan, Director of Psychological Services, Lansing, Michigan, for his support and assistance.

[2] At the beginning of the study there were 62 different employers and 84 possible *Ss*. Thirteen of the possible 84 *Ss* could not be included in the study: 2 dropped out of the program and could not be tested, 3 had unusable rating sheets because of misunderstood instructions, and 8 employers failed to return rating sheets.

were in a work-training program for the educationally handicapped in operation at the high school level in the Lansing, Michigan, Public Schools. Students in this program go to school on a part-time basis and work part time for employers cooperating in the program. The Special Education Department maintains classes in three high schools in the city of Lansing, and every *S* attends that school which is geographically closest to his home. Each *S* in the program had obtained a WISC, WAIS, or Stanford-Binet IQ score between 50 and 90, had a history of course work failure for the last two or more years, and showed retardation of two or more years on scholastic achievement tests. Participation in the program was initially recommended by the appropriate school authority, required the concurrence of the parent(s), and was voluntarily entered into by the student.

Cutting Score (CS) and Replication (R) Samples. The 40 *S*s who were in the work-training program during the first year it was fully instituted formed the CS sample. The 31 *S*s who were in the work-training program during its second year of existence formed the R sample.

Measures

Porteus Maze Test (PMT). The PMT was administered and scored for each *S* in accordance with the instructions set forth by Porteus (1959). The *S*'s derived test age score was used for purposes of analysis. Following Porteus, this score was based on the total number of mazes successfully completed, taking into consideration the number of trials required.

Rating Scale of Personal Effectiveness (RSPE). Personal effectiveness was assessed by a rating scale devised by Warren (1961) for use with mentally retarded adolescents. The scale is made up of 18 items, the items having to do with such characteristics as punctuality, cooperativeness, and showing initiative.[3]

A RSPE was filled out on each *S* by both his teacher and employer. Neither evaluator knew that the other one was also filling out a rating scale on the *S*. The teacher was asked to rate the *S* on

---

[3] The relationship between scores on the PMT and each item on the RSPE was determined in a pilot study. There was a high positive relationship between the PMT and 10 of the items, a low positive relationship between the PMT and 5 of the items, and no relationship between the PMT and 3 of the items.

each item in relation to his average student in the program, and the employer in relation to his average employee. For each item the *S* was to be rated as "less than," "same as," or "more than" the average student or employee.[4]

In order to avoid obvious contamination of the criterion measure, only the teachers' ratings were scored for use in data analysis. The RSPE was scored to yield a total weighted rating score for each *S*. For any item, "less than" ratings were assigned a weighted score of 1, "same as" a score of 2, and "more than" a score of 3. The minimum possible score was 18 and the maximum 54.

Employability

At the bottom of the RSPE sheet, employers checked one of four categories—Yes, Probably, Probably Not, No—in response to the question, "Would you be willing to hire this individual as you would your average applicant, if a job were available?" This employability item was the criterion measure. Those *S*s for whom the employer checked either "Yes" or "Probably" were classed together under Employable. Those *S*s for whom the employer checked either "Probably Not" or "No" were classed together under Not Employable. In the case of the R sample, it was also possible to obtain the additional criterion measure of who actually did or did not get hired later by the employers.

Procedure

The Porteus test was administered to each *S* in a private office of the school he was attending.

At the head of each group of Special Education students in the three high schools is a teacher. Besides teaching he has a regular time set aside each day for counseling and placement. He obtains work-training employment for the students with an employer willing to cooperate in the program, and maintains close communication with the employer in order to gauge the students' work progress. The three teachers were asked to fill out a RSPE on each of their student-trainees. They were given a form letter which described the project and gave instructions on how to fill in the rating sheet. In the form letter they were informed that their responses would be used in a research project and they were encouraged to be as accu-

[4] In the pilot study the teacher and employer ratings correlated .94.

rate as possible. They were also reassured that their ratings would in no way influence the *S*'s school or job standing.

Every employer was sent a rating sheet with his trainee's name on it, a copy of another form letter, and a stamped addressed envelope for the return of the rating sheet. The form letter sent to the employer was the same as that given to the teacher-coordinator except for slight changes of phrasing to make it more appropriate to the work situation.

The steps in data collection were the same for all *S*s. During the time the rating scales were being filled out and returned by the teachers and employers, the PMT was administered and scored for all the *S*s. After all the rating sheets were returned, or otherwise accounted for, they were scored. The scored protocols for both the PMT and RSPE were then identified as belonging to a particular *S* in either the CS or R groups. The data were subsequently assembled and analyzed separately for first the CS and then the R sample.

Results

There were no sex or age differences on either the PMT or RSPE. Therefore, sex and age were disregarded in all analyses. Fisher exact probability tests and one-tailed rejection regions were used for all comparisons. The p values reported are exact probabilities. The entries in the cells in Tables 1 and 2 are number of *S*s.

The results for the CS sample are most conveniently considered first; then the results for the R sample.

CS Sample

PMT. A test age cutting score of 13.7 years on the PMT gave maximum success in discrimination. Part A of Table 1 presents the relevant summary data. A Fisher exact probability test yields a p value of .007. The employers indicate they would hire 78 per cent of the *S*s in this sample, and although correct placement does not occur for 100 per cent of the cases, it is impressively frequent. On the basis of their PMT score 32 out of 40 *S*s are correctly placed with respect to whether they would be hired. Inspection of the data in Part A also shows that the hit-rate is relatively better within the Employable than it is within the Not Employable cate-

gory; or, in other words, the proportion of false negatives within the Employable category is less than the proportion of false positives within the Not Employable category.

RSPE. A cutting score of 33.5 on the RSPE was maximally effective in discriminating those who were likely to be hired or not. The probability associated with a set of observations as or more extreme than those in Part B of Table 1 is .0005. Correct placement is again impressively frequent; 35 out of 40 *S*s are correctly classified. In comparison to the corresponding PMT data, the proportion of false negatives is about the same while the proportion of false positives drops a bit.

PMT and RSPE. A brief re-examination of Parts A and B shows that better prediction is obtained for those *S*s scoring above than for those scoring below the cutting score on either the PMT or RSPE. Both score distributions also show some negative skewing, although this is not apparent from the data presented. In Part C of Table 1 the data have been assembled using the *S*'s score position on both the PMT and RSPE.

Inspection shows that using this double standard results in perfect placement of those *S*s scoring above or below the cutting score on both scales. It eliminates both the false positives and false negatives. A Fisher test based on the data presented in the upper portion of Part C gives a p value of .00001. The set of differences observed in the lower portion is not statistically significant, indi-

TABLE 1

Comparison of Score Position and Employer's Judgment for CS Subjects

Part	Test	Score[a] Position	Judgment of Employer: Employable	Judgment of Employer: Not Employable
A	PMT	Plus	26	3
		Minus	5	6
B	RSPE	Plus	27	1
		Minus	4	8
C	PMT–RSPE	Plus–Plus	22	0
		Minus–Minus	0	5
		Plus–Minus	4	3
		Minus–Plus	5	1

[a] Plus indicates above the cutting score, minus below.

cating that scoring above the cutting value on the RSPE and below on the PMT is not significantly more likely to result in being considered employable than scoring above on the PMT and below on the RSPE.

R Sample

PMT. Having established cutting scores on the PMT and RSPE with the CS sample, the next question was whether or not these scores would continue to be as predictive with a new sample of *Ss*.

The cutting score on the PMT derived with the CS group continued to give a high rate of correct prediction. Inspection of Part A of Table 2 shows that correct prediction was obtained for 23 of the 31 *Ss*, and this overall hit-rate is comparable to that observed in the CS sample. A Fisher test on the set of observations in Part A gives a p value of .015. The employers indicate that they would hire 61 per cent of the *Ss* in this sample, a slight drop in the proportion from that observed in the CS sample. Although perfect placement of these *Ss* does not occur, it is relatively frequent; 79 per cent are correctly placed. A further statistical check on the per-category hit-rate differences for the CS and R samples shows that they do not differ significantly.

RSPE. The original cutting value on the RSPE also continued to yield a high rate of correct prediction. The probability associated

TABLE 2

Comparison of Score Position and Employer's Judgment for R Subjects

Part	Test	Score[a] Position	Judgment of Employer: Employable	Judgment of Employer: Not Employable
A	PMT	Plus	15	4
		Minus	4	8
B	RSPE	Plus	16	0
		Minus	3	12
C	PMT–RSPE	Plus–Plus	13	0
		Minus–Minus	1	8
		Plus–Minus	2	4
		Minus–Plus	3	0

[a] Plus indicates above the cutting score, minus below.

with a set of observations as or more extreme than those in Part B is .000003. Twenty-eight of the total 31 *S*s are correctly placed. There are no false positives, and there are three false negatives. The overall and per-category hit-rates observed here differ very little from those observed with the CS sample (see Table 1, Part B).

Somewhat higher overall and per-category hit-rates with the RSPE than with the PMT again occur with the R sample. However, the overall and—except in one instance—the per-category hit-rate differences of the two scales are not statistically significant for either the CS or the R sample. The one exception occurs with the R sample where the hit rate in the Employable category is significantly higher using the RSPE ($p < .01$, by binomial expansion).

PMT and RSPE. Further inspection of the data indicated that, as with the CS sample, better prediction is obtained in those instances where a *S* scores above or below the critical value on both scales. In Part C of Table 2 the data have been assembled using the *S*'s score position on both the PMT and RSPE. The probability associated with a set of observed values as or more extreme than those in the upper portion of Part C is .00003. Perfect prediction is again obtained for those *S*s scoring above the critical values. Except for one false negative, this would also be the case for those *S*s scoring below the critical values.

As with the CS sample it also appears from inspection that *S*s scoring above the cutting score on the RSPE and below on the PMT are more likely to be judged hirable by the employer. However, the set of differences presented in the lower portion of Table 2 are not significant. In addition, if the data from both samples are combined, the differences are still not significant.

Finally, as previously noted, it was possible to find out who among the 31 *S*s in the R sample eventually got hired and who did not. Of the 12 *S*s classified as Not Employable, none was hired later by his respective employer. Of the 19 *S*s classified as Employable, 18 were hired later by their respective employers. The remaining *S* would have been hired, but his family had moved out of the state.

Discussion

As Special Education Departments continue to expand their school work-training programs, and as the number of potential em-

ployers who can be enlisted to cooperate in such programs approaches an asymptote, more careful screening of those to be trained will become increasingly necessary. Given this eventuality, the Porteus Maze test could prove to be a helpful screening device. Apart from the fact that it is relatively inexpensive as well as simple and quick to administer, the present results suggest that it might profitably be used as an initial screening instrument. If, for example, in the present study only those *S*s scoring above the cutting score had been admitted into the training program, 85 per cent of them would presumably have been hired by their employers.

Should this relatively high rate of initial screening be judged not high enough, however, the results for the rating scale indicate that if it is used in combination with the Porteus Maze test, even further successful screening might be possible. If, for instance, only those *S*s who scored *above* the critical value on *both* scales had been admitted into the training program, 100 per cent of them would presumably have been hired by their employers.

Data were not collected that would allow some kind of quantification of savings—in terms of cost, time, or efficiency of training in the program—to be derived by screening out those who would later prove to be unhirable. It does seem likely, however, that an appreciable and worthwhile savings could accrue by instituting both scales as screening procedures. The Porteus Maze test could be used to select those who will initially enter the program, and the teachers' ratings could be obtained early in the training program and used for final selection. Assuming the school and work-training program involved in the present study to be representative of both the situation where there are and the situation where there are not test screening procedures in use, the instance where the tests are in use would seem to produce the greatest savings. Although 30 per cent of those who would be considered hirable would be screened out of the program by using both scales, this loss might be more than offset by the savings that would accrue from screening out all (100 per cent) of those who would presumably not be hired.

It remains to be seen, of course, just how specific to the conditions, *S*s, and program setting the present results are. Further research may or may not show they have some more general applicability.

REFERENCES

Porteus, S. D. *The Maze Test and Clinical Psychology*. Palo Alto: Pacific Books, 1959.

Warren, F. G. Ratings of Employed and Unemployed Mentally Handicapped Males on Personality and Work Factors. *American Journal of Mental Deficiency*, 1961, 65, 629–633.

Educational and Psychological Measurement
1966, 26, 1031–1034.

PREDICTING SUCCESS IN A VOCATIONAL REHABILITATION PROGRAM WITH THE RAVEN COLOURED PROGRESSIVE MATRICES

KENT L. KILBURN and ROBERT E. SANDERSON
Porterville State Hospital
California Department of Mental Hygiene

There has been interest in the prediction of vocational success by psychological tests of intelligence for several years as evidenced by the work of Shafter (1957) and Walker (1951). The use of other types of psychological tests in this capacity has been explored by Arnholter (1962), Ferguson (1958), Kolstoe (1961), Novis, Marra, and Zadrozny (1960), Taylor (1963), and Warren (1961). Madison (1964) and Tobias and Gorelick (1963) have shown a positive relationship between intellectual level and vocational success. More specifically Appell, Williams, and Fishell (1962), Fry (1956), and Jackson and Butler (1963) have noted the utility of the Wechsler Scale in this area. However, Reger and Dawson (1961) indicated that a Wechsler score did not adequately predict occupational therapy success, and Parnicky and Kahn (1963) found no relationship between Peabody Picture Vocabulary scores and vocational success. Aside from the limited success noted for Wechsler scores, the prediction of vocational success with psychological tests, particularly intellectual tests appears to have not been determined.

This paper is concerned with assessing the degree to which the Raven Coloured Progressive Matrices (RCPM) (Raven 1956a) and the Peabody Picture Vocabulary Test (PPVT) (Dunn 1959) predict patient success in a vocational rehabilitation program (VR) at a hospital for the mentally retarded. When the VR program was reorganized in June of 1964, selection of each of the 40 patients

for the program was made primarily on the criteria of his past activities, current behavior, and general potential based on a subjective evaluation of the patient's speech and socioadaptive behavior level.

Independently of this selection process, all patients who were considered for the VR program were tested with the RCPM, PPVT, and the Stanford-Binet word vocabulary. The staff of the VR program had no access to the RCPM or PPVT test results. Thus, their decisions on patient disposition in the program were made without knowledge of these data.

During the 18 months that the VR program has been in operation, 12 patients have been placed successfully into a work situation outside the hospital, and 9 patients have been dropped from the program as failures because of the lack of adequate work performance or displays of asocial behavior as determined by the VR program staff. As these patients were moved out of the program, new ones were added to keep the number at approximately 40.

Methods and Results

The group of 21 patients were split into the 12 successes and the 9 failures. A t test for the difference between mean scores and point biserial correlation were computed between each sub-group for each test in order to determine the power of the raw scores of the RCPM and PPVT to predict success in the VR program.

The results shown in Table 1 indicated a difference significant beyond the .05 level between the RCPM mean score of 16.46 for success, and 12.56 for failure. The point biserial correlation be-

TABLE 1

Distribution of Scores for VR Patients

		Mean	SD	*N*	*Range*	*t for Mean Difference*	*rpb*
RCPM	Success	16.46	4.22	12	11–24		
	Failures	12.56	2.24	9	9–16	2.73*	.60**
PPVT	Success	70.58	14.83	12	53–106		
	Failures	64.11	10.18	9	51–83	1.18	.27

* Significant beyond the .05 level.
** Significant beyond the .01 level.

tween success or failure and RCPM scores of .60 was significant beyond the .01 level.

However, the difference between the PPVT mean score of 70.58 for success and 64.11 for failure was not significant. Similarly, the point biserial correlation between success or failure and the PPVT scores of .27 was not significant.

Discussion

The lack of a significant relationship between PPVT scores and success or failure in the VR program is similar to the finding of Parniky and Kahn (1963) who indicated that a test of verbal intelligence does not relate to vocational placement. The significant positive relationship between RCPM scores and status in a vocational training program is supported by the findings of Appell, Williams, and Fishell (1962) who found Wechsler scores to be useful in predicting vocational training success or failure. The RCPM and Wechsler both showing positive relationships to VR success seems reasonable, as Orme (1961) found a positive correlation of .93 between the Wechsler Adult Scale and the RCPM in a sample of 203 retarded subjects. The RCPM described by Raven (1956b) as "a test of observation and clear thinking" indicates an important aspect of success in a VR program. Success in such a training program involving actual prevocational experiences seems to be more related to the ability to observe and to solve new problems than merely to recall word meanings. Although an interesting trend has been observed, caution should be applied in generalizing the present results. This trend should be re-examined with larger groups of various types of VR training programs.

REFERENCES

Appell, M. J., Williams, C. M., and Fishell, K. N. Significant Factors in Placing Mental Retardates from a Workshop Situation. *The Personnel and Guidance Journal*, 1962, 41, 260–265.

Arnholter, E. G. The Validity of Fisher's Maladjustment and Rigidity Scales as an Indicator of Rehabilitation. *The Personnel and Guidance Journal*, 1962, 40, 634–637.

Dunn, L. M. *Peabody Picture Vocabulary Test*. Minneapolis: American Guidance Service, Inc., 1959.

Ferguson, R. G. Evaluating Vocational Apptitudes and Characteristics of Mentally Retarded Young Adults in an Industrial-Agricultural Workshop. *American Journal of Mental Deficiency*, 1958, 62, 787–791.

Fry, Lois M. A Predictive Measure of Work Success for High Grade Mental Defectives. *American Journal of Mental Deficiency*, 1956, 61, 402–408.

Jackson, Sue K. and Butler, A. J. Prediction of Successful Community Placement of Institutionalized Retardates. *American Journal of Mental Deficiency*, 1963, 68, 211–217.

Kolstoe, O. P. An Examination of Some Characteristics Which Discriminate between Employed and Not-Employed Mentally Retarded Males. *American Journal of Mental Deficiency*, 1961, 66, 472–482.

Madison, H. L. Work Placement Success for the Mentally Retarded. *American Journal of Mental Deficiency*, 1964, 69, 50–53.

Novis, F. W., Marra, J. L., and Zadrozny, L. J. Quantitative Measurement in the Initial Screening of Rehabilitation Potential. *The Personnel and Guidance Journal*, 1960, 39, 262–269.

Orme, J. E. The Coloured Progressive Matrices as a Measure of Intellectual Subnormality. *British Journal of Medical Psychology*, 1961, 34, 291–292.

Parnicky, J. J. and Kahn, H. *Evaluating and Developing Vocational Potential of Institutionalized Retarded Adolescents*. Vocational Rehabilitation Administration Grant #425, United States Department of Health, Education, and Welfare at Edward R. Johnstone Training and Research Center, Bordentown, New Jersey, 1963.

Raven, J. C. *Coloured Progressive Matrices (1947) Sets A, Ab, B, Book Form*. London: Lewis, revised, 1956. (a)

Raven, J. C. *Guide to Using the Coloured Progressive Matrices (1947), Sets A, Ab, B, Revised*. London: Lewis, 1956. (b)

Reger, R. and Dawson, Antoinette. The Use of Psychological Tests to Predict Manual Abilities in Mentally Retarded Boys. *American Journal of Occupational Therapy*, 1961, 15, 204–221.

Shafter, A. J. Criteria for Selecting Institutionalized Mental Defective for Vocational Placement. *American Journal of Mental Deficiency*, 1957, 61, 599–616.

Taylor, F. R. The General Appitude Test Battery as Predictor of Vocational Readjustment by Psychiatric Patients. *Journal of Clinical Psychology*, 1963, 19, 130.

Tobias, J. and Gorelick, J. Work Characteristics of Retarded Adults at Trainable Levels. *Mental Retardation*, 1963, 1, 338–344.

Walker, J. L. Psychological Tests as Predictors of Vocational Adjustment. *American Journal of Mental Deficiency*, 1951, 56, 429–432.

Warren, F. G. Ratings of Employed and Unemployed Mentally Handicapped Males on Personality and Work Factors. *American Journal of Mental Deficiency*, 1961, 65, 629–633.

Educational and Psychological Measurement
1966, 26, 1035–1040.

THE PREDICTIVE VALIDITIES OF SELECTED APTITUDE AND ACHIEVEMENT MEASURES AND OF THREE PERSONALITY INVENTORIES IN RELATION TO NURSING TRAINING CRITERIA

WILLIAM B. MICHAEL
University of California, Santa Barbara
RUSSELL HANEY and ROBERT A. JONES
University of Southern California

Problem

For a sample of one hundred freshman trainees in student nursing for the academic year of 1964–1965, the two-fold purpose of this investigation was (1) to obtain additional cross-validation data on a number of cognitive and non-cognitive predictor variables that had been employed during previous years in the selection of candidates for participation in a nursing training program at the Los Angeles County Hospital and (2) to derive new information concerning the predictive validities of scales in the Edwards Personal Preference Schedule (EPPS) and in Cattell's Sixteen Personality Factor (16 PF) Questionnaire relative to both academic and clinically oriented criteria. In addition, it was also desired to check out the predictive validities of the previously employed Minnesota Multiphasic Personality Inventory (MMPI) scales with respect to certain items on the Ward Performance Scale that had been designed to represent operational statements of several of the constructs in the EPPS and in the 16 PF Questionnaire. The constructs selected were judged to reflect personal needs thought to be important by supervisory staff to success in day-to-day nursing activities in the wards of any large metropolitan hospital.

Predictor and Criterion Variables

In Table 1 the ten predictor variables representing measures of scholastic aptitude and achievement are enumerated along with the

fourteen criterion measures employed. Only those noncognitive predictors (subscales of two personality inventories—the EPPS and the 16 PF Questionnaire) are cited that showed statistically significant (.05 level) predictive validity coefficients with three or more of the criterion measures. (The MMPI scales are not listed because none of them satisfied the statistical requirement just defined.)

Statistical Treatment

With the exception of the predictor represented by grades in high school chemistry which were placed on a four-point scale, all measures were converted to stanine scores by a normalized-rank method (Haney, Michael, and Gershon, 1962). Through use of an IBM 7090 program at the Western Data Processing Center at UCLA, product-moment coefficients of correlation were calculated among all possible pairings of variables. Since there was a marked restriction of range as evidenced by the fact that the standard deviations of scores for the total group of several hundred applicants were from 25 to 80 per cent larger than those of the survival group, the coefficients reported in Table 1 are minimal in size. An example of such restriction is apparent from the fact that on the total scores of the California Reading Test and on the total scores of the California Mathematics Test trainees were customarily required to place at least above the 50th centile with respect to published 12th grade norms.

Results

The following major findings may be summarized for the data in Table 1:

1. For the academic criteria the part and total scores of the California Reading Test, and the Reasoning and total scores of the California Mathematics Test showed for the most part statistically significant coefficients of predictive validity. However, neither of these instruments appeared to be predictive of rated characteristics of performance in the wards.

2. With the exception of the criterion variable of grades in Psychology II, the somewhat speeded measure of spatial ability—EAST No. 5—yielded one positive and twelve negative coefficients which failed to attain statistical significance.

TABLE 1

Predictive Validity Coefficients of Measures for the 1964–1965 Nursing Class at the Los Angeles County Hospital ($N = 100$)

Predictors (Variables 1–14) and Criterion Measures (Variables 15–28)	Validity of Coefficients of Predictors Along with Intercorrelations of Criterion Measures* (All Decimal Points Omitted)													
	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)**	(28)**
1. Calif. Reading Test (Vocabulary)	26	21	30	18	10	35	28	39	23	27	17	29	17	08
2. Calif. Reading Test (Comprehension)	20	22	28	20	24	16	21	34	23	18	10	23	−02	08
3. Calif. Reading Test (Total)	25	28	35	22	20	32	30	48	25	33	24	36	14	11
4. Calif. Math Test (Reasoning)	37	22	29	16	12	09	18	25	24	17	16	10	03	10
5. Calif. Math Test (Fundamentals)	34	10	17	01	04	−02	07	09	05	−01	02	01	12	02
6. Calif. Math Test (Total)	34	21	32	11	12	03	18	26	19	16	18	13	13	11
7. EAST No. 5—Space Visualization	−01	−17	−07	−17	−10	−16	−10	−12	−28	−12	07	−11	−06	−01
8. Edwards PPS—Order	22	14	32	17	08	29	28	18	17	22	12	02	02	00
9. Edwards PPS—Autonomy	−18	−23	−25	−14	−21	−07	−17	−26	−14	−13	−03	−20	−05	03
10. 16 PF Questionnaire—L	00	−31	−25	−22	02	05	−08	−26	−16	−10	−17	−19	−10	−05
11. 16 PF Questionnaire—O	−22	−14	−11	−22	−02	−05	−08	−16	−20	−07	−13	−12	−05	−01
12. High School GPA—All Solids	34	42	37	42	20	41	38	41	26	48	27	35	18	11
13. High School GPA—All Courses	39	37	40	41	30	39	36	40	22	47	35	33	20	18
14. High School Chemistry (Two Semester GPA)	30	26	32	36	09	26	31	27	21	42	20	23	09	09
15. Arithmetic Grades	—	33	37	33	23	34	45	11	16	31	20	28	24	16
16. Anatomy Grades	33	—	52	56	27	30	45	59	43	63	32	46	21	09
17. Nursing I Grades	37	52	—	59	44	37	51	62	41	51	43	63	29	09
18. Nutrition Grades	33	56	59	—	36	45	54	59	50	53	36	56	20	20
19. Orientation I Grades	23	27	44	36	—	34	33	41	29	34	35	41	15	18
20. Orientation II Grades	34	30	37	45	34	—	55	35	32	50	25	45	18	16
21. Microbiology Grades	45	45	51	54	33	55	—	54	52	58	24	50	18	14
22. Psychology I Grades	11	59	62	59	41	35	54	—	66	55	33	59	14	14
23. Psychology II Grades	16	43	41	50	29	32	52	66	—	50	29	47	08	08
24. Physiology Grades	31	63	51	53	34	50	58	55	50	—	32	54	20	04
25. Surgical Nursing, Clinical Grades	20	32	43	36	35	25	24	33	29	32	—	52	34	48
26. Surgical Nursing, Examination Marks	28	46	63	56	41	45	50	59	47	54	52	—	37	30
27. Ward Performance Scale (May 1965)	24	21	29	20	15	18	18	14	08	20	34	37	—	37
28. Ward Performance Scale (June 1965)	16	09	09	20	18	16	14	14	08	04	48	30	37	—

* Uncorrected for restriction of range coefficients of .20 and .26 are significant at the .05 and .01 level.

** The values are based on the median coefficient of each of the predictor variables (1–14) and of the 14 criterion measures (15–28) with the 12 items of the Rating Scale of Ward Performance.

3. As in previous studies (e.g. Haney, Michael, and Gershon, 1962; Michael, Haney and Gershon, 1963; Michael, Haney, and Brown, 1965) high school grade point average (GPA) whether in all subjects or in academic subjects alone served as the best overall predictor of success in the academic portion of the training program. In the current sample, the GPA in high school chemistry did not afford a single predictive validity coefficient that was not surpassed by the high school GPA in all courses. Moreover, none of the GPA variables was even moderately predictive of success in ward performance.

4. Although failing to be predictive of success in ward performance, two scales in the EPPS—order and autonomy—and two scales in the 16 PF Questionnaire—L and O—did show statistically significant degrees of relationship with at least three academic criterion measures. It would appear from a study of the size and sign of coefficients for the two EPPS scales that relatively high standings in the trait order and relatively low placement in the trait autonomy were slightly predictive of success in the academic program. Designating a trustful (adaptable) versus a suspecting (jealous) and a confident (unshakable) versus an insecure (anxious) orientation, the L and O scales, respectively, in the 16 PF Questionnaire yielded for this sample validity coefficients which suggested that for at least three of the academic criteria trainees who were trustful and confident were more likely to succeed in the academic program than were those who were suspicious and insecure in their adjustment to academic requirements.

5. As in previous studies with nursing trainees at the Los Angeles County Hospital the degree of correlation among grades in academic courses was usually higher than that found between cognitive predictors and grades in these same courses. As was pointed out elsewhere (Michael, Haney, and Gershon, 1963), the presence of a moderate degree of correlation among course marks (usually between .30 and .60) suggested one or more of the following hypotheses: (a) a "halo" effect in grading practices, (b) a motivational syndrome related to earning grades, (c) an academic savoir-faire factor embodying a sensitivity of students to the needs and expectations of their teachers, and (d) an institutional press in which students respond to imposed standards of behavior in a generally conforming manner.

6. Although the presence of moderate coefficients of correlation (which are not reported here, but customarily were found between .30 and .70) among several of the pairs of the twelve items of the Ward Performance Scale at both the May and June evaluations might indicate that the characteristics were intercorrelated to some degree, the likelihood of a general halo effect in the rating process could not be discounted despite the care with which instructions on rating procedures had been communicated to the supervisory staff. The median test-retest coefficient of correlation of each of the twelve characteristics evaluated in May with each of the eleven other characteristics and with itself evaluated in June was .37. For the May evaluation and for the June evaluation the median coefficients of correlation of every one of the twelve characteristics with each of the other characteristics (but not with itself) were .54 and .48, respectively. (Application of the Spearman-Brown formula yielded internal consistency estimates in reliability of .94 and .92 for the May and June evaluations.)

7. Despite the fact that each of the twelve characteristics in the Ward Performance Scale was developed to measure activities that were judged coordinate with constructs in the EPPS and to some extent with those in the 16 PF Questionnaire the coefficients of correlation observed between measures of the constructs in the inventories and the corresponding characteristics evaluated in the Ward Performance Scale hovered near zero and approached significant values at only a chance level (i.e., for about five per cent of the entries obtained). Thus neither the scholastic aptitude and achievement measures nor the scales in the MMPI, in the EPPS, or in the 16 PF Questionnaire were predictive of the measures of performance in the ward. It should also be noted that even the median coefficients of correlation of each of the twelve characteristics on the Ward Performance Scale with each of the other criterion measures in the academic program approximated values that were either barely significant at the .05 level or only close to significance.

Conclusion

For a sample of one hundred freshmen in student nursing who survived one year of training at the Los Angeles County Hospital moderate and statistically significant coefficients of validity for measures of high school GPA, reading comprehension, and mathe-

matics reasoning were found relative to academic courses in the program. For the most part scales of the MMPI, EPPS, and the 16 PF Questionnaire showed a lack of validity either for academic course work in nursing training or for rated performance in ward activities. The presence of moderate to high intercorrelations among marks in the academic courses as well as among items on the Ward Performance Scale suggested not only the probable presence of a halo effect but also a motivational syndrome for earning high marks in both academic and clinical activities.

REFERENCES

Haney, R., Michael, W. B., and Gershon, A. Achievement, Aptitude, and Personality Measures as Predictors of Success in Nursing Training. Educational and Psychological Measurement, 1962, 22, 389–392.

Michael, W. B., Haney, R., and Gershon, A. Intellective and Non-intellective Predictors of Success in Nursing Training. Educational and Psychological Measurement, 1963, 23, 817–821.

Michael, W. B., Haney, R., and Brown, Stephen, W. The Predictive Validity of a Battery of Diversified Measures Relative to Success in Student Nursing. Educational and Psychological Measurement, 1965, 25, 579–584.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 1041–1046.

A PLACEMENT STUDY IN ANALYTIC GEOMETRY AND CALCULUS[1]

RICHARD L. FRANCIS
Southeast Missouri State College

It is the purpose of this article to present certain portions of a recent placement study in mathematics at the University of Missouri. In particular, an effort was made to assess the values of selected test scores for the prediction of freshman success in the first and second courses of a three semester sequence in analytic geometry and calculus. Since the study was concerned with freshmen, the third course was not considered. These two integrated courses in analytic geometry and calculus are ordinarily referred to as Calculus I and Calculus II.

As to pre-requisites, a grade of C or better was required in Trigonometry and College Algebra in order to qualify for Calculus I. Otherwise one had to be a select student with 3.5 to 4 high school units of mathematics including two units of algebra, one unit of geometry, and one-half unit of trigonometry. As calculus seemingly demands proficiency in algebraic and trigonometric concepts and manipulation, it was necessary that any uncertainty here as to qualifications be resolved. Also the high number of deficient grades received in Calculus I and Calculus II gave evidence of the need for a valid basis for assigning pupils to these courses.

Method of Analysis

Scores on the Mathematics Advanced Placement Tests (Algebra and Trigonometry) were taken as two of the variables in the regression analysis for the prediction of Calculus I achievement. These

[1] The writer wishes to thank Dr. H. W. Schooling and Dr. F. G. Delon, his advisors in this study, for their most valuable help and suggestions.

tests were prepared by the Missouri Section of the Mathematical Association of America and are given annually at the university for the purpose of advanced placement. This is a three hour examination with 130 points possible on the algebra section and 70 points possible on the trigonometry section. One additional predictor variable was used: the School and College Ability Test quantitative score. This test is administered to all freshmen having less than 24 semester hours of credit; thus these scores were available for the entire experimental group. There were 50 points possible on the quantitative examination.

The construction of the test instrument to measure achievement in Calculus I was one of the initial steps in the experiment. In order to ensure comprehensiveness, a survey was conducted involving the teachers of calculus in the department of mathematics. Also, test papers of the instructors were examined to determine the frequency and types of errors. The information obtained was used as a basis for determining the content of the achievement examination.

For the final form of the achievement test, the thirty most valid items were selected from the sixty items initially constructed. The discriminating power of each item was determined by administering the first form of the examination to a select group of students. By considering the proportion of correct responses in the upper and lower subgroups, the validity index for each item was read directly from a table of normalized biserial coefficients. The method of rational equivalence (Richardson and Kuder, 1939) was used to provide a measure of reliability. This coefficient was obtained by using the scores of the forty people in the pilot group and was found to be .769 with the standard deviation equal to 4.892. To further substantiate the reliability of the examination, a coefficient was calculated based on the 569 students taking the Calculus I Achievement Examination in January of 1965. Here the reliability coefficient was found to be .778 with the standard deviation equal to 4.60.

Findings of the Calculus I Study

Scores on the Mathematics Advanced Placement Tests were available for only 254 students of the freshman group enrolling in Calculus I. SCAT-Q scores were available for all of these however.

By least squares analysis using the score weights for Algebra (X_2), Trigonometry (X_3), and Quantitative (X_4) the various regression equations of Table 1 resulted for the prediction of Calculus I achievement (X_1) scores. The coefficient of multiple correlation between the scores obtained on the criterion variable of Calculus I achievement (X_1) and the combined action of the independent variables X_2, X_3, and X_4 was .479. The $SE_{(est.\ X1)}$ was equal to 4.073.

The best single predictor of Calculus I achievement (X_1) was the Algebra score (X_2). The correlation between the criterion and this measure of mathematical ability was .402 (significant at the .01 level). The various intercorrelations are listed in Table 2.

The point-biserial coefficient of correlation was calculated for the pass-fail dichotomies of teacher assigned marks in Calculus I and learning achievement in Calculus I as measured by the Calculus I Achievement Examination. The value 1 was assigned to a passing grade (C or better) in the course and the value 0 was assigned to a failing grade, with the test criterion being the score on the achievement examination. This correlation was found to be .526 which is significant at the .01 level. The calculation of this coefficient seemed important, since the study was limited by the teacher variable as it pertained to assigned marks in Calculus I and Calculus II.

Findings of the Calculus II Study

The size of this second experimental group was 143 and consisted of those from the Calculus I study who enrolled in Calculus II.

TABLE 1

Regression Equations and Standard Errors of Estimate of Calculus I Achievement Scores for a Combination of the Variables Algebra, Trigonometry, and Quantitative ($N = 254$)

Variables Combined	Regression Equations	$SE_{(est.\ X_1)}$
Algebra (X_2) and Trigonometry (X_3)	$\hat{X}_1 = 16.4806 + .07996X_2 + .03288X_3$	±4.236
Algebra (X_2) and Quantitative (X_4)	$\hat{X}_1 = 4.57174 + .06415X_2 + .29764X_4$	±4.097
Trigonometry (X_3) and Quantitative (X_4)	$\hat{X}_1 = 1.19767 + .08082X_3 + .41199X_4$	±4.196
Algebra (X_2), Trigonometry (X_3), and Quantitative (X_4)	$\hat{X}_1 = 3.73188 + .05482X_2 + .04274X_3 + .30767X_4$	±4.073

TABLE 2

Intercorrelations between a Measure of Achievement in Calculus I and Selected Measures of Mathematical Ability ($N = 254$)

Test	Coefficients of Correlation			
	X_2	X_3	X_4	X_1
Algebra (X_2)	—	.3880	.3929	.4015
Trigonometry (X_3)	—	—	.0789	.2228
Quantitative (X_4)	—	—	—	.3814
Calculus I (X_1)	—	—	—	—

Regression equations involved the numerical equivalent of the semester grade received in the course, called the performance or criterion, and was denoted by Y. The letter grades A, B, C, D, and F were assigned the respective values 4, 3, 2, 1, and 0. The Calculus I achievement, algebra, trigonometry, and quantitative variables from the previous study became the predictor variables. The various regression equations are listed in Table 3. The coefficient of multiple correlation between the scores obtained on the criterion variable of Calculus II achievement (Y) and the combined action of the independent variables X_1, X_2, X_3, and X_4 was .607. The $SE_{(\text{est. } Y)}$ was equal to .901.

The best single predictor of Calculus II achievement was the Calculus I (X_1) score. Here r was equal to .521 (significant at the .01 level). The various intercorrelations are listed as Table 4.

Concluding Remarks

Since more than half of the variability of achievement in Calculus I and Calculus II was not explained by the measures considered in this study, further investigation is suggested. Answers to the following questions could perhaps result from this further investigation:

1. What specific study habits are associated with a high level of achievement in Calculus I and Calculus II?
2. Is the effect of such qualities as interest, motivation, and stability in the learning of calculus quantitative and measurable?
3. The findings of this study indicate a need for more valid placement instruments. What procedures are most effective

TABLE 3

Regression Equations and Standard Errors of Estimate of Calculus II Achievement Scores for a Combination of the Variables Calculus I Achievement, Algebra, Trigonometry, and Quantitative (N = 143)

Variables Combined	Regression Equations	$SE_{(est.\ Y)}$
Calculus I (X_1) and and Algebra (X_2)	$\hat{Y} = -1.54713 + .12990X_1 + .01589X_2$	$\pm .90063$
Calculus I (X_1) and Trigonometry (X_3)	$\hat{Y} = -1.73085 + .16100X_1 + .01772X_3$	$\pm .94012$
Calculus I (X_1) and Quantitative (X_4)	$\hat{Y} = -2.66591 + .16438X_1 + .02692X_4$	$\pm .95705$
Algebra (X_2) and Trigonometry (X_3)	$\hat{Y} = .89947 + .02137X_2 + .01135X_3$	$\pm .97996$
Algebra (X_2) and Quantitative (X_4)	$\hat{Y} = .35757 + .02249X_2 + .01580X_4$	$\pm .98624$
Trigonometry (X_3) and Quantitative (X_4)	$\hat{Y} = -1.22866 + .02558X_3 + .06648X_4$	± 1.05551
Calculus I (X_1), Algebra (X_2), and Trigonometry (X_3)	$\hat{Y} = -1.61559 + .12809X_1 + .01430X_2 + .00927X_3$	$\pm .89823$
Calculus I (X_1), Algebra (X_2), and Quantitative (X_4)	$\hat{Y} = -1.41362 + .13048X_1 + .01606X_2 - .00342X_4$	$\pm .90379$
Calculus I (X_1), Trigonometry (X_3), and Quantitative (X_4)	$\hat{Y} = -2.78128 + .15278X_1 + .01772X_3 + .02689X_4$	$\pm .93882$
Algebra (X_2), Trigonometry (X_3), and Quantitative (X_4)	$\hat{Y} = .06238 + .02008X_2 + .01194X_3 + .01959X_4$	$\pm .98130$
Calculus I (X_1), Algebra (X_2), Trigonometry (X_3), and Quantitative (X_4)	$\hat{Y} = -1.61053 + .12811X_1 + .01431X_2 + .00927X_3 - .00012X_4$	$\pm .90148$

TABLE 4

Intercorrelations between a Measure of Learning Achievement in Calculus II and Selected Measures of Mathematical Ability and Achievement (N = 143)

Test	Coefficients of Correlations				
	X_1	X_2	X_3	X_4	Y
Calculus I (X_1)	—	.3954	.1913	.2863	.5206
Algebra (X_2)	—	—	.3620	.3856	.4807
Trigonometry (X_3)	—	—	—	.0553	.2758
Quantitative (X_4)	—	—	—	—	.2286
Calculus II (Y)	—	—	—	—	—

for the construction and evaluation of other freshman placement tests in mathematics?

4. What is the relative value of high school analytic geometry and calculus in the determination of placement and how do these courses relate to achievement in college calculus?
5. What is the value of the Mathematics Advanced Placement Tests in the prediction of achievement in mathematics courses other than Calculus I and Calculus II?

Because of the complexities involved in the learning of calculus, many factors should be considered when placing a student. Certain of the variables mentioned above do not lend themselves to easy measurement. Hence, if they are to be considered, it must be in a subjective way by the teacher and the advisor. The use of mathematical ability and achievement tests in the various manners described in this study to supplement the judgment of teachers and advisors should contribute to the more accurate placement of students in Calculus I and Calculus II.

REFERENCES

Garrett, H. E. *Statistics in Psychology and Education.* New York: David McKay Company, 1962.

Richardson, M. W. and Kuder, G. F. The Calculation of Test Reliability Coefficients Based upon the Method of Rational Equivalence. *Journal of Educational Psychology,* 1939, 30, 681–687.

Educational and Psychological Measurement
1966, 26, 1047–1052.

THE RELIABILITY AND CORRELATES OF AN ACHIEVEMENT INDEX

ERICH P. PRIEN
The University of Akron
AND
DAVID E. BOTWIN
University of Pittsburgh

Several methods are currently used to quantify an achievement index (AI). The AI is defined here as the difference between predicted performance and actual performance. Actual performance may be predicted by a single measure of ability, or by using a combination of aptitude and ability measures. Graphic calculation of deviations from a regression line (or computations) yields a continuous score distribution, just as does a difference between standard scores. Various other techniques have been used, all of which are intended to identify *relative* achievement criterion groups.

The appropriate method of analysis, to study over-achievement and under-achievement according to Thorndike (1963), is a partial correlation coefficient without the computation of a separate AI. However, deviations were calculated in this study to permit computation of the reliability or stability of relative achievement over time. The AI was then used for correlations with personality test scores rather than the part correlation technique.

The primary purpose of this study was to investigate the reliability of an AI over a two year period. The secondary purpose was to determine the correlation of the AI with selected personality characteristics.

Method

Subjects were the complete junior class of 101 students in a small, private women's college. Complete data were obtained from ninety-five students.

Measures

The criterion data available for this study were the grade point averages for the freshman year (GPA-1), the sophomore year (GPA-2), the fall term of the junior year (GPA-3), and the *cumulative* performance for the first two years (GPA 1-2).

All students had initiated their college career at the sponsoring institution and had taken a battery of aptitude and achievement tests. A selection process based primarily on high school graduating class rank and to some extent on the Verbal (V) and Mathematics (M) portions of the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board was in effect at the time of admission. However, some latitude was allowed the admissions authority to accommodate individual cases where past circumstances indicated a reasonable probability of success. Thus, while some restriction in range is probable for predictor scores, the effect is assumed to be minimal. Selective attrition, however, is more likely relevant to the problem of restriction of range.

The personality measures used in this study were the Gordon Personal Profile, (Gordon, 1963b), Gordon Personal Inventory (Gordon, 1963a), and the Survey of Interpersonal Values, (Gordon, 1961). These specific tests were selected by the authors because of personal experience with the measures in research and in extensive student counseling. Personality tests were administered to the students early in the Fall of their junior year. Administration was on a group basis, with assurance given to the students by the college administration that their performance on this test would not be used in any administrative action as far as their group was concerned. Apparent complete cooperation was obtained from the group.

The AI was computed separately for each year and for the two year cumulative GPA. The Ohio State University Psychological Examination (OSUPE) was used as the predictor. The multiple correlation obtained using the OSUPE in combination with any of the other aptitude measures did not materially increase the predicta-

TABLE 1

Product-Moment Correlations, Means and Standard Deviations of Academic Aptitude Measures and Criterion Measures (N = 95)

Measures	1	2	3	4	5	6	7	8	Mean	Sigma
1. SAT Verbal		45	75	73	48	41	35	46	473.	89.0
2. SAT Math			51	52	41	44	32	45	470.	76.4
3. OSUPE				64	60	44	37	51	93.8	22.1
4. COOP Reading					53	45	30	53	60.8	8.2
5. GPA-1						76	61	92	2.51	.54
6. GPA-2							65	92	2.62	.50
7. GPA-3								65	2.67	.65
8. GPA-1, 2 (Combined)									2.57	.48
$R_{8.3,1} = .52$										
$R_{8.3,2} = .56$										
$R_{8.3,4} = .57$										

Note.—Decimal points omitted from correlation coefficients only.
* All correlations are significant at the .01 level.

bility of the criteria. The multiple correlations (R) computed to predict GPA 1, 2 when the OSUPE and SAT verbal were used was: .52; when the OSUPE and SAT mathematics were employed, R was .56; when OSUPE and COOP reading C-2 were chosen, R was .57. The question of using a multiple regression equation to calculate the AI was resolved after consideration of the point in time of administration of each test. The SAT scores were obtained some time prior to entrance, the OSUPE at the beginning of the freshman year, and the COOP Reading C-2 at various times during the freshman year or at the beginning of the sophomore year. Since there is already a question of temporal separation to be considered in the interpretation, further confounding seemed undesirable. The intercorrelation of aptitude test scores with the individual year and cumulative GPA is shown in Table 1.

In this study, relative achievement was a continuum, and the entire range of the variable was used instead of employing extreme groups as had Keimowitz and Ansbacher (1960), Krug (1959), and Diener (1960). Computation of the AI was done graphically from the plotted regression line. Accuracy was judged by the authors as approximately $\pm .02$, of a grade point. Thus, if the predicted GPA was 2.55 and the actual GPA was 2.89, the AI would be $.34 \pm .02$. A small sample of calculations was made and compared with the much faster graphic approach and the accuracy was deemed satisfactory for present purposes. Negative values were eliminated from the AI by the addition of a constant.

Results

The significant correlations between the AI computed separately for the first year, second year, and first term of the third year, indicated the existence of true variation in the GPA which is independent of academic aptitude. These data appear in Table 2.

Product-moment correlation coefficients which were computed between personality measures and GPA 1, 2 and with the AI 1, 2

TABLE 2

Intercorrelations and Standard Deviations of Individual Achievement Indices Calculated for the Freshman, Sophomore, and First Term of the Junior Year. (N = 95)

Year	2	3	Sigma
1	.70*	.58*	.44
2		.57*	.45
3			.59

Note.—All correlations are significant at the .01 level.
* Significant at the .05 level.

TABLE 3

Correlations, Means, and Standard Deviations of GPA 1, 2, and AI, and Personality Variables (N = 95)

Variable	OSUPE	GPA(1,2)	AI(1,2)	Mean	Sigma
Gordon Personal Profile					
Ascendancy	05	−05	−07	3.59	8.4
Responsibility	−04	17	28**	3.73	6.1
Emotional Stability	−10	06	21*	2.89	6.7
Sociability	01	−19	−19	4.55	6.8
Gordon Personal Inventory					
Cautiousness	−06	27**	43**	1.58	6.4
Original Thought	27**	36**	25*	6.00	6.4
Personal Relations	−20*	08	24*	4.06	6.4
Vigor	03	17	14	4.21	5.8
Survey of Personal Values					
Support	−02	00	08	17.00	4.8
Conformity	−27**	−23*	−06	11.35	5.7
Recognition	−01	−03	00	10.57	4.3
Independence	04	10	05	20.49	6.2
Benevolence	−13	−01	13	16.47	6.3
Leadership	11	06	−05	12.07	6.2
Mean	93.78	2.57	.00		
Sigma	22.08	.48	.39		

Note.—Decimal points omitted from correlations only.
* Significant at the .01 level.
** Significant at the .05 level.

appear in Table 3. As is readily apparent, there is considerable similarity of correlations with the GPA 1, 2 and the AI 1, 2. Correlations which appear between the personality variables with GPA 1, 2 and with the AI 1, 2 cannot be construed as representing causal relationships. Rather, it can only be said that the significant cluster of traits, in either case, describes students who tend to receive relatively higher or lower grades than do their peers.

With the contribution of academic aptitude eliminated, relative achievement is significantly related to personality characteristics. The cluster of variables shows positive correlations with cautiousness, responsibility, personal relations, emotional stability, and negative correlation with sociability. The cluster suggests that relatively high achieving students behave differently or at least says that they do behave differently than do low relative achievers. We may speculate that the high achievers are actually more productive, more accurate, more painstaking, and thus, better students who deserve relatively higher grades, rather than beneficiaries of rater halo.

Generally, the results obtained in this study are similar to those found by Krug (1959) who described the overachiever as scoring higher than underachievers on the Edwards Personal Preference Schedule (Edwards, 1959) subscales Achievement, Order, Endurance; and lower on the subscales of Affiliation and Heterosexuality. The behavioral description offered by Krug corresponds to that espoused in this study.

Finally, whether one subscribes to the interpretation of instructor halo, or of the actual fulfillment of student role behaviors resulting in real superiority of performance, the relationship between personality factors and relative achievement is significant.

REFERENCES

Diener, D. L. Similarities and Differences between Over-achieving and Under-achieving Students. *Personnel and Guidance Journal*, 1960, 38, 396–400.

Edwards, A. F. *Edwards Personal Preference Schedule Manual.* New York: Psychological Corp. 1959.

Gordon, L. V. *Survey of Interpersonal Values Manual.* Chicago: Science Research Associates, 1961.

Gordon, L. V. *Gordon Personal Inventory Manual.* New York: Harcourt, Brace, and World, Inc., 1963. (a)

Gordon, L. V. *Gordon Personal Profile Manual.* New York: Harcourt, Brace, and World, Inc., 1963. (b)

Keimowitz, R. I. and Ansbacher, H. I. Personality and Achievement in Mathematics. *Journal of Individual Psychology*, 1960, 16, 84–87.

Krug, R. E. Over and Under-achievement and the Edwards Personal Preference Scale. *Journal of Applied Psychology*, 1959, 43, 133–136.

Thorndike, R. L. *The Concepts of Over- and Under-achievement.* New York: Bureau of Publications, Teachers College, Columbia University, 1963.

Educational and Psychological Measurement
1966, 26, 1053–1054.

PERSONALITY AND GRADES OF COLLEGE STUDENTS OF DIFFERENT CLASS RANKS[1]

RICHARD M. SUINN
University of Hawaii

Studies of the relationship between personality and college grades have used various types of questionnaires and projective tests. Of such, the Guilford-Zimmerman Temperament Survey (Guilford and Zimmerman, 1949) should be of greatest value since it measures traits more appropriate for describing the normal college student. The purpose of this study was to examine the relationship between Guilford-Zimmerman scores and grades on upper and lower division students.

Method

Results from the Guilford-Zimmerman test were obtained on 184 private liberal arts college students: 47 freshmen, 46 sophomores, 46 juniors, and 45 seniors. Grade-point averages for the last semester were obtained. For the freshmen, grade predictions were also available. These latter were based on weighted Verbal and Math scores of the Scholastic Aptitude Test of the College Entrance Examination Board (CEEB) and high school grade-point average from previous studies of incoming freshmen.

Results

Grade-point average was significantly correlated with seriousness for freshmen ($r = .43, p < .01$), juniors ($r = .41, p < .01$), and all students combined ($r = .27, p < .01$); and with friendliness for sophomores ($r = .32, p < .05$), seniors ($r = .33, p < .05$), and all students combined ($r = .21, p < .01$).

[1] This study was supported in part from a grant from the National Institute of Mental Health, grant no. M.R. 07490-01.

For the freshmen men and women, grade *prediction* scores were subtracted from their actual grade achievements to give deviation scores for use in correlations. Since the grade prediction formula included CEEB scores, this method controlled for differences in scholastic aptitude. Results showed that seriousness and the degree to which a student performed better than predicted were significantly related ($r = .44$ for men; .43 for women, $p < .05$). Thus, students described as serious or as showing restraint tended to achieve grades higher than had been expected of them from their high school record and college aptitude scores.

REFERENCE

Guilford, J. P. and Zimmerman, W. S. *The Guilford-Zimmerman Temperament Survey*. Calif.: Sheridan Supply Co., 1949.

Educational and Psychological Measurement
1966, 26, 1055–1056.

OTIS PREDICTION OF GRADUATE EDUCATION COURSE GRADES

A. M. FOX and L. L. AINSWORTH
Sam Houston State College

The purpose of this study was to determine the extent to which Otis Quick-Scoring Mental Abilities Test (OQS) IQ's may be useful for the prediction of grade-point ratios (GPR's) in graduate Education courses.

Sample

The sample consisted of 306 students at Sam Houston State College for whom undergraduate OQS IQ's and at least one grade in a graduate Education course were available, for the period July, 1959, through August, 1964.

Procedure

GPR's were calculated, and then relationships of OQS IQ's to GPR's were determined by the Stepwise Regression Analysis Program (STRAP) in an IBM 1620 Data Processing System. An F-ratio of 4.0 was used as the criterion in testing the significance of the relationships. Significant relationships were found for OQS IQ's with total GPR (GPRT), basic required courses (GPRB), Measurement and Evaluation (GPR93), Secondary Curriculum (GPR94), Human Growth and Development (GPR97), superintendent's certificate courses (GPRS), supervisor's certificate courses (GPRV), and principal's certificate courses (GPRP). Regression equations, coefficients of alienation (k), and indices of forecasting efficiency (E) were then computed for these significant relationships.

Results

The results of this study are summarized in Table 1.

TABLE 1

Relationships of OQS to Dependent Variables

Dependent Variable	N	r	k	OQS IQ		GPR		Regression Equation	σyx	E
				Mean	S.D.	Mean	S.D.			
GPRT	306	0.33	0.94	111.6	9.7	3.2	0.53	2.2 + 0.018 (IQ)	0.50	6%
GPRB	281	0.32	0.94	111.7	9.6	3.1	0.59	1.9 + 0.020 (IQ)	0.56	6%
GPR93	235	0.31	0.95	112.1	9.7	3.2	0.68	1.8 + 0.021 (IQ)	0.65	5%
GPR94	148	0.26	0.97	112.4	9.2	3.2	0.66	2.1 + 0.018 (IQ)	0.64	3%
GPR97	148	0.37	0.93	112.4	9.0	3.0	0.70	0.8 + 0.029 (IQ)	0.65	7%
GPRS	35	0.37	0.93	111.5	8.7	3.3	0.67	1.1 + 0.029 (IQ)	0.54	7%
GPRV	66	0.32	0.94	110.4	8.6	3.2	0.49	2.2 + 0.018 (IQ)	0.48	6%
GPRP	102	0.35	0.94	110.7	8.8	3.1	0.50	1.9 + 0.020 (IQ)	0.47	6%

Educational and Psychological Measurement
1966, 26, 1057-1062.

CONCURRENT VALIDITY OF THE GATES LEVEL OF COMPREHENSION TEST AND THE BOND, CLYMER, HOYT READING DIAGNOSTIC TESTS

DONALD A. BENZ
Wisconsin State University
at Stevens Point

AND

ROBERT A. ROSEMIER
Northern Illinois University[1]

The purpose of this study was to investigate the relationship between word-analysis skill proficiency to reading comprehension performance among fourth graders. This attempt differs from many prior studies in that it did not structure the sample according to intelligence or other factors and thereby restrict this investigation of the relationship to "over" or "under-achievers." Rather, it allowed a determination among these variables on children "as they were." It should contribute additional information concerning the "substrata model" of the reading task as described by Holmes and Singer (1964).

Variables

The six word-analysis skills selected for investigation were: Syllabication, location of root words, words in context, word elements, beginning sounds, and rhyming sounds, as defined by the Bond, Clymer, Hoyt (1955) Silent Reading Diagnostic Tests, Form D-A. Reading comprehension, serving as the criterion variable, was assessed by the Gates (1958) Level of Comprehension Test, Type LC, Form 3.

[1] Presently with the Southwest Regional Laboratory for Educational Research and Development, Los Angeles.

Subjects

The participants consisted of all fourth grade children attending public schools in six towns and cities in five northeastern states. These fifty-five classrooms of 1490 children included what the authors believed to be a representative cross-section of school environments; included were a college town (N =273), a resort village (N = 148), a state capital (N = 115), an industrial-professional community (N = 274), an agricultural center (N = 248), and a so-called "bedroom" community serving a nearby metropolitan center (N = 432). Although none of the communities was considered "culturally deprived," a wide range of environments did exist among and within the systems considered. The number of classrooms in the systems ranged from five to eighteen. The population appeared to be only slightly above normal (mean = 107) on the basis of intelligence scores which were available for 94 per cent of the group from a variety of tests.

Procedures

Classroom teachers were trained by the investigators to administer the tests which were given in consecutive half-day testing sessions during a one month interval. The tests were scored by the investigators. Of the 1490 fourth-grade children tested, those scoring between 6.0 and 9.2 on the Gates test were considered high level readers (N = 474), those between 4.6 and 5.9 as middle level (N = 450), and those from 2.0 to 4.5 as low level (N = 478). A number of subjects were dropped from future analyses either because of absence during testing or because of failure to achieve the minimum level reported on the comprehension test. As a result, 1402 children were available for analysis.

Analyses

The analyses which were performed resulted from the following lines of reasoning. If a skill was contributory to reading comprehension, it should discriminate among various proficiency levels of reading comprehension. Further, it should show a high degree of correlation with reading comprehension when the confounding effects of related variables are removed or "partialed out." When taken as a whole, these six word analysis skills should account for much of the existing error variance in reading comprehension.

The first approach required an analysis of variance and subsequent Scheffé tests (Hays, 1963, pp. 483–485) among three levels of student reading comprehension for each of the six word-analysis skills. The second involved the computation of a partial correlation coefficient between reading comprehension scores and scores on a given skill, while the effects of the other five skills were held constant (or removed from the correlation), for each of the six skills (Ezekiel and Fox, 1959, pp. 192–197). The third interest involved a step-wise multiple regression analysis (Smillie, 1962).

Results

The results of the first phase indicated that each of the six word-analysis skills did discriminate among the various levels of reading proficiency. The mean scores and the standard deviations on each skill for each of the three reading comprehension levels are shown in Table 1.

TABLE 1

Mean-scores and Standard Deviations Attained by High, Middle, and Low Reading Comprehension Levels for Six Word-Analysis Skills

Skill	High		Middle		Low	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
Words in context (1)	25.46	2.015	23.22	2.640	20.58	3.521
Syllabication (2)	15.98	3.503	13.15	3.517	10.70	3.363
Root Word (3)	23.85	3.730	22.07	3.994	19.09	4.496
Word Elements (4)	26.69	2.423	24.93	2.916	22.59	3.716
Beginning Sounds (5)	27.55	2.713	26.16	3.373	23.91	4.473
Rhyming Sounds (6)	25.64	3.360	22.66	4.326	19.14	4.672
	$N=474$		$N=450$		$N=478$	

The summary table for the simple randomized analyses of variance are presented in Table 2.

These analyses and the ensuing tests of differences between the various pairs of levels, which were carried out according to procedures developed by Scheffé and described in Hays (1963, pp. 483–485), indicated the higher comprehension readers performed better than middle comprehension readers who performed better than low comprehension readers on each word analysis skill at the .01 level of significance.

For phase two, partial correlation coefficients were computed between reading comprehension scores and scores on a given word

analysis skill while the effects of the remaining five word analysis skills were statistically removed. These coefficients are shown in Table 3.

The standard-score regression weights for the six word-analysis skills resultant in the last phase are also shown in Table 3, as are the standard errors and t-values for testing their significance in the regression analysis. The intercorrelation matrix among the six word-analysis skills and reading comprehension employed in this multiple regression analysis is shown in Table 4. When these word-

TABLE 2

Summary Table for Analysis of Variance among Levels (L) by Skills

Source		*df*	*ms*	*F*	
Skill	Between L's	2	2,849.392	364.70	*
1	Within L's	1,399	7.813		
	Total	1,401			
Skill	Between L's	2	3,321.956	277.34	*
2	Within L's	1,399	11.978		
	Total	1,401			
Skill	Between L's	2	2,747.657	164.75	*
3	Within L's	1,399	16.678		
	Total	1,401			
Skill	Between L's	2	2,012.561	207.47	*
4	Within L's	1,399	9.701		
	Total	1,401			
Skill	Between L's	2	1,629.822	126.10	*
5	Within L's	1,399	12.925		
	Total	1,401			
Skill	Between L's	2	5,032.852	292.01	*
6	Within L's	1,399	17.235		
	Total	1,401			

* Significant at .01 level. ($F = 4.61$, $df = 2,1399$).

TABLE 3

Partial Correlation Coefficients, Beta (β) Weights, Standard Errors of Beta-weights, and t-values for Testing Significance of Six Word-Analysis Skills

Skill	Beta-weight	σ_β	t	Partial r's
Words in context	.318903	.0168	18.99**	.5548
Syllabication	.182966	.0144	12.75**	.4300
Root Word	.110184	.0124	8.89**	.3230
Word Elements	.067276	.0095	7.08**	.2651
Beginning Sounds	.010058	.0041	2.47*	.0975
Rhyming Sounds	.229511	.0154	14.95**	.4793

* Significant at .05 level ($t = 1.96$, $df = \infty$).
** Significant at .01 level ($t = 2.58$, $df = \infty$).

TABLE 4

Intercorrelation of Skill Scores and Reading Comprehension Scores (N = 1,402)

Test		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Words in Context	(1)		.544	.497	.517	.438	.565	.642
Syllabication	(2)			.480	.541	.449	.531	.572
Root Words	(3)				.433	.337	.425	.486
Word Elements	(4)					.596	.573	.516
Beginning Sounds	(5)						.551	.435
Rhyming Sounds	(6)							.598
Reading Comprehension	(7)							

analysis skills were considered in the analysis, the multiple correlation coefficient was .734—a finding which indicated that composite, optimally-weighted tests accounted for approximately 54 per cent of the variation in reading comprehension.

A subsidiary check was performed on these last two analyses by also considering the variable of intelligence. It should be noted that although intelligence scores were available for 94 per cent of the sample, these scores were attained from a variety of intelligence tests. Hence any interpretation which is made from the addition of this variable is, at best, a rough estimate of its contribution.

The partial correlations of each of the six word analysis skills and reading comprehension mentioned earlier differed by no more than .02 when intelligence was added. The multiple regression coefficient was raised only to .753 and the percentage of variation accounted for by the six measures of skills with the addition of intelligence variable was increased to approximately 57 per cent.

Conclusions

Proficiency on the six word analysis skills seemed to be significantly related to reading comprehension performance of fourth graders on the basis of the discrimination among various levels of reading comprehension on each skill. Five of the six skills (omitting the skill of "beginning sounds") appeared to contribute significantly on the basis of their partial correlation coefficients, as was also observed from tests of significance of the beta weights within the regression analysis. Approximately 46 per cent of the variance in the measure of reading comprehension was *not* ac-

counted for by the six word-analysis skills and must be considered attributable to extraneous variables not under consideration.

REFERENCES

Bond, Guy L., Clymer, Theodore, and Hoyt, Cyril. *Silent Reading Diagnostic Tests, Form D-A*. Chicago: Lyons and Carnahan, 1955.

Ezekiel, Mordecai and Fox, Karl A. *Methods of Correlation and Regression Analysis—Linear and Curvilinear*. New York: John Wiley & Sons, 1959.

Gates, Arthur I. *Gates Basic Reading Survey for Grades 3.5–10*. New York: Bureau of Publications, Teachers College, Columbia University, 1958.

Hays, William L. *Statistics for Psychologists*. New York: Holt, Rinehart and Winston, 1963.

Holmes, Jack A. and Singer, Harry. Theoretical Models and Trends toward More Basic Research in Reading. *Review of Educational Research*, 1964, 34, 127–155.

Smillie, K. W. Program No. 1620-002, Stepwise Regression. Ottawa, Canada: Dominion Bureau of Statistics, 1962. (mimeographed)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 1063–1067.

STABILITY OF MMPI SCALES OVER FIVE TESTINGS WITHIN A ONE-MONTH PERIOD[1, 2]

JEROME D. PAUKER
Department of Psychiatry
University of Missouri Medical School

THE Minnesota Multiphasic Personality Inventory (MMPI) (Hathaway and McKinley, 1951) is frequently used in before-and-after studies as a measure of change resulting from the manipulation of some independent variable (as, for example, in studies of drug effects). The MMPI changes to be expected concommitant with the passage of time (without systematic insertion of intervening events) have been demonstrated in the various control groups used in experimental studies and in the results of retest reliability studies (Dahlstrom and Welsh, 1960).

This paper presents the MMPI data of ten young women, each of whom took the MMPI *five* times during one month as part of a study of the menstrual cycle. The purpose of this investigation is to show how each of the "validity" and "clinical" MMPI scales (plus the Welsh A and R scales) fares in the course of such repeated testing.

Procedure

Ten women, college students, volunteered to be subjects in a study of the psychological course of the menstrual cycle. They were required to be between 18 and 24 years of age, to have regular menstrual cycles of between 27 and 29 days duration, to be unmarried, and not to be taking oral contraceptives or other drugs which might affect the menstrual cycle. Seven of the ten subjects were stu-

[1] This study was supported by a grant from the Medical and Cancer Research Fund of the University of Minnesota Graduate School.
[2] The suggestions of Dr. Starke R. Hathaway and of Dr. Auke Tellegen are acknowledged with appreciation

dent nurses in a class taught by the writer. The seven were all in the first year of nursing school and in the second year of college. The other three subjects were recruited from a sorority to which one of the student nurses belonged. The mean age was 20.2 years, with a range of from 19 to 21 years.

All subjects had taken the MMPI on entering the university, and the seven nursing students had taken it again about two months before the start of the present testing as part of another study.

Information was obtained from each subject about the duration of her menstrual cycle and the point in the cycle at which she then was. Arrangements were then made for each subject to take the MMPI on the second day of menstruation, the eighth day in her menstrual cycle, the fifteenth day in the cycle, the twenty-second day in the cycle, and two days before menstruation (Test A, B, C, D, and E, respectively).

This schedule was varied a bit depending upon the actual length of the cycle and to avoid testing on Sundays. Each subject took her first MMPI (Test 1) at the first test point in her menstrual cycle which came up subsequent to the interview. One subject took her first MMPI at point A (see schedule above), two took their first MMPI at point B, three at point C, one at point D, and three at point E. For the group, Test 1 and Test 2 were separated by a mean of 5.3 days (range 3–7 days), Test 2 and Test 3 by a mean of 5.5 days (range 4–8 days), Test 3 and Test 4 by a mean of 5.5 days (range 2–9 days), and Test 4 and Test 5 by a mean of 5.4 days (range 3–7 days).

An IBM card form of the MMPI was used and was scored by an IBM 101. In a few instances (as when a subject had to be away from school on a test day), the booklet form was used.

In order to promote the most open response to the test, each subject was given a code number; the number was the only identification which went on each test, and the test results were filed until the end of the study at which time each subject was paid fifteen dollars and a card bearing the subject's name and code number was destroyed.

Results

Table 1 presents the mean T-scores (K-corrected) of each scale for each testing (from the first test, regardless of point in menstrual

cycle, to the fifth test). For most of the scales, the change in mean T-score from testing to testing is under 3.0 points; in only three instances is this exceeded (an increase of 4.2 on scale 3 from Test 2 to Test 3, a decrease of 3.8 on scale 4 from Test 1 to Test 2, and an increase of 5.6 on scale 5 fromTest 2 to Test 3). The median change is 1.2 T-score points.

Analysis of variance in a treatments by subjects design (Lindquist, 1953) was used to test the hypothesis that there are no differences between the groups. Using the mean squares for testings ($df = 4$) and for treatments by subjects ($df = 36$) as the numerators and denominators, respectively, the F-ratios listed in Table 2 were computed. None of the F-ratios was significant at the .05 level.

Discussion

The results of the analyses of variance show that the observed changes in mean T-scores from testing to testing could be expected to occur often enough on the basis of chance that one would want to be cautious in attributing them to any systematically intervening factor. The mean changes in themselves, even taken at face value, are not large.

The reason for lack of change is not readily apparent from this study. It could be due to a lack of significant change in the subjects, this subject stability being accurately reflected in the MMPI

TABLE 1

Mean T-Scores (Ten Subjects) of 15 MMPI Scales for Each of Five Testings

	Test 1	Test 2	Test 3	Test 4	Test 5
Scale L	44.9	45.0	45.1	45.2	44.0
Scale F	51.7	49.9	50.4	48.7	49.9
Scale K	54.0	55.0	57.0	57.1	57.8
Scale 1 (*Hs*)	48.6	47.9	48.0	46.9	48.0
Scale 2 (*D*)	50.0	48.0	46.0	46.6	46.2
Scale 3 (*Hy*)	57.0	57.6	53.4	53.8	56.1
Scale 4 (*Pd*)	56.5	52.7	53.9	51.3	53.0
Scale 5 (*Mf*)	41.8	41.9	47.5	45.0	43.5
Scale 6 (*Pa*)	56.6	58.9	56.8	59.0	58.5
Scale 7 (*Pt*)	55.7	55.5	56.7	54.6	55.8
Scale 8 (*Sc*)	58.2	55.3	58.2	55.5	56.5
Scale 9 (*Ma*)	62.6	61.2	61.9	61.8	61.3
Scale 0 (*Si*)	49.1	49.7	47.2	48.5	47.2
Scale A	50.7	51.4	49.3	47.9	46.2
Scale R	48.8	48.5	48.6	49.3	47.5

Note.—For those unacquainted with the MMPI, a brief description of the test and of each scale may be found in Hathaway and McKinley (1951). The A and R scales are described in Dahlstrom and Welsh (1960).

TABLE 2

F-Ratios Resulting from Analyses of Variance of T-Scores (K-Corrected) of 15 MMPI Scales Administered Five Times to Ten Subjects

MMPI Scale	F-Ratio[a]	MMPI Scale	F-Ratio[a]
L	0.37	6 (Pa)	0.92
F	1.07	7 (Pt)	0.25
K	1.23	8 (Sc)	0.87
1 (Hs)	0.20	9 (Ma)	0.17
2 (D)	1.12	0 (Si)	1.33
3 (Hy)	1.84	A	2.40
4 (Pd)	1.29	R	0.42
5 (Mf)	1.83		

[a] None of the F-ratios is significant at the .05 level (F of 2.69 required for 5 per cent level with 4 and 30 degrees of freedom).

results. It could also be due to a lack of sensitivity to change, either because the MMPI does not do the job it should do (psychologists at Minnesota, would possibly resist such a consideration) or because the short intervals between testings permitted memory to play a role in determining responses to the items.

Measures of the stability of individual MMPI scales and two-point codes (Dahlstrom and Welsh, 1960) and of MMPI profiles (Pauker, 1965, in press) have demonstrated respectable levels of retest reliability. The present study shows that repeated testing with the MMPI, at least up to five testings and with a young female college group, has no marked effect on group scale means.

Summary

Ten young women, college students, took the MMPI five times during a one-month period. The data were grouped for each scale for each testing. Analysis of variance was done with the *K*-corrected *T*-scores of each of the three "validity" scales, the ten "clinical" scales, and the A and R scales. None of the resulting *F*-ratios was significant at the .05 level.

REFERENCES

Dahlstrom, W. G. and Welsh, G. S. *An MMPI Handbook.* Minneapolis: The University of Minnesota Press, 1960.

Hathaway, S. R. and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory Manual. Revised.* New York: The Psychological Corporation, 1951.

Lindquist, E. F. *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton Mifflin Co., 1953.

Pauker, Jerome D. MMPI Profile Stability in a Psychiatric Inpatient Population. *Journal of Clinical Psychology*, 1965, 21, 281–282.

Pauker, Jerome D. Stability of MMPI Profiles of Female Psychiatric Inpatients. *Journal of Clinical Psychology*, 1966, 22, 209–212.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 1069–1072.

SHIFTS IN MEASURES OF ATTITUDES OF MEDICAL STUDENTS TOWARD THOSE OF THEIR PROFESSORS RELATIVE TO THE DOCTOR IMAGE AND THE DOCTOR-PATIENT RELATIONSHIP: IMPLICATIONS FOR PREDICTION OF A CLINICALLY ORIENTED CRITERION

SEYMOUR POLLACK[1] AND WILLIAM B. MICHAEL
University of Southern California School of Medicine
AND
University of California, Santa Barbara

In two previous studies, the writers (Pollack and Michael, 1965; 1966) investigated changes in the attitudes of medical students toward psychological characteristics of the doctor image and the doctor-patient relationship on four empirically derived factor scales in Blum's (1957) Patient Attitude Test (PAT)—three scales—and in the Doctor's Opinion Questionnaire (DOQ)—one scale. Although only the mean changes on the dimension concerned with fee-setting practices could be judged as statistically significant, there were, at that time, no available data concerning the attitudes of doctors as measured on the three dimensions of the PAT, against which the attitudes of medical students could be compared.

Purpose

It was the purpose of this exploratory investigation to determine the extent and direction of shift in the standing of medical students on each of twelve tentatively identified attitudinal dimensions concerned with the doctor image and the doctor-patient relationship

[1] This research was supported by Research Grant No. MH-07366-01 from the National Institute of Mental Health to the University of Southern California School of Medicine.

toward the composite standing on the same dimensions of thirty-one clinical and thirty-one full time professors of a university medical school—dimensions which consisted of items from the PAT and the DOQ. The importance of such a pilot study was thought to lie in the tentative description of the nature of attitudes of medical students and in the determination of possible changes in the admittedly crude measures of those non-cognitive components that are seemingly inherent in any criterion of success in medical school. In other words, it appeared that in the prediction and evaluation of success of medical school students—especially in clinical activities—any substantial shift of measurable levels of their attitudes in a direction toward those levels held by their professors could be of practical significance to a student's success in clinical practice in that such success may depend, in large measure, upon a student's assimilating and accepting, or at least of giving the impression of assimilating and accepting, the attitudinal and value systems held by his professors.

Procedure

Items of the PAT and DOQ which met a criterion of endorsement by at least seventy-five percent of both the thirty-one clinical and thirty-one teaching staff at the University of Southern California Medical School (see Pollack and Michael (1966) for additional details) were subjected to a preliminary content analysis involving four judges from whose pooled judgments emerged twelve identifiable categories that are set forth primarily for their heuristic value in future scale construction. For samples of 59 freshmen, 55 sophomores, 40 juniors, and 33 seniors, who in early October, 1961 responded to PAT and DOQ items placed in a questionnaire, the percentages of their endorsements were calculated and compared with the standards established for the 62 professors.

Results

In Table 1, the twelve *a priori* categories are cited, along with median percentages of endorsements of items in each category, by members of each of the four classes and by the sample of physician-teachers. Although the highly exploratory and preliminary nature of this investigation did not appear to justify the application of significance tests in the absence of cross-validation data, and de-

TABLE 1

Median Percentages of Endorsement for Items in Each of Twelve Response Categories Concerned with the Doctor Image and Doctor-Patient Relationship

Response Category	Number of Items[a]	School Class				Doctors (Teaching and Clinical Staff)
		Fresh.	Soph.	Jrs.	Snrs.	
1. Rendering medical services fairly	6	74	80	75	78	88
2. Feeling of competence in practice	13	87	89	85	91	95
3. Fee-setting procedures	13	78	73	80	85	87
4. Doctor as a person	5	83	91	91	91	86
5. Medicine as a calling	2	84	79	86	85	83
6. Professional ethics	6	71	74	73	78	91
7. Doctor's liability	5	73	71	75	86	89
8. Organic (pathological) vs. patient orientation	23	73	78	78	85	84
9. Self-confidence in medical practice	11	50	58	68	68	96
10. Feelings of satisfaction and ease in dealing with patient	4	77	84	91	91	87
11. Consideration of patient's psychological needs	5	96	90	86	94	98
12. Socializing with the patient	3	39	38	50	58	81

[a] Twenty-eight of the 95 items appeared in two or more categories.

spite limitations which must be posed in terms of the possible operation of response sets of acquiescence and social desirability in the respondents answering the question, inspection of the median percentages tentatively suggests not only that the perceptions of medical students in each class are similar to those of the professors, but also that, in general, there are small shifts from year to year in a direction indicating a slightly closer correspondence of the perceptions of students to those of their professors. Although the findings reveal that there is a rather close similarity between the preliminary measures of the attitudes of medical students and those of their professors—a correspondence which seemingly becomes slightly greater as students progress in their program—additional investigations involving numerous cross-validation efforts are needed before reliable and valid scales can be developed, not only to predict those noncognitive components for the clinically oriented criterion that

were suggested by this pilot study but also to exhibit changes in the perceptions of medical students relative to those held by their professors.

REFERENCES

Blum, R. H. *The Psychology of Malpractice Suits.* Prepared from the Medical Review and Advisory Board of the California Medical Association, San Francisco, March, 1957.

Pollack, S. and Michael, W. B. Changes in Attitudes of Medical Students Toward Psychological Aspects of the Doctor-Image and the Doctor-Patient Relationship. *Journal of Medical Education,* 1965, 40, 1162–1165.

Pollack, S. and Michael, W. B. Prelimnary Normative Data Regarding Attitudes of Clinical and Teaching Staffs of a University Medical School Toward Psychological Characteristics of the Doctor-Image and the Doctor-Patient Relationship. Unpublished manuscript. 1966.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1966, 26, 1075–1119.

BOOK REVIEWS

WILLIAM B. MICHAEL, Editor
University of California, Santa Barbara
JOAN J. BJELKE, Assistant Editor
Centinela Valley Union High School District
and the
University of Southern California

Statistical Analysis in Psychology and Education by George A. Ferguson. New York: McGraw-Hill Book Co. 1966. Pp. x + 446.

This second edition of Ferguson's popular text follows the first edition by seven years. (See Binder's review of the first edition in the Winter 1960 issue of this journal.) A considerable amount of new material has been included (approximately 60 pages exclusive of exercises); relatively little of the content of the first edition was revised. Those who were devotees of the 1959 edition of *Statistical Analysis in Psychology and Education* will probably find much to be pleased with in the 1966 edition. This competently done and comprehensive edition seems likely to capture an even wider audience than did the first.

The first eight chapters of the first and second editions of *Statistical Analysis in Psychology and Education* are nearly identical. From Chapter 1 "Basic Ideas in Statistics" through the hard-core material on graphs, tables, measures of central tendency and variability, probability, and the normal curve to Chapter 8 "Prediction in Relation to Correlation," little has been changed. To be sure, some sections have been polished (the chapter on probability has been slightly improved), some have been eliminated (the geometric and harmonic means have disappeared and that vestige of a bygone era, the bivariate table for the calculation of r, has been mercifully laid to rest), and the exercises, which were generally poor in the first edition, are now passable. However, the bulk of the first 130 pages of both the first and second editions are about the same. Ferguson handles this elementary material competently, but without particular distinction. The only substantial change from the first edition is in the definition of the sample variance. In the first edition, s^2 was the biased, maximum likelihood estimator of σ^2; in this edition, the unbiased estimator of σ^2 is used. This is clearly an improvement. Ferguson appears to find the unbiased estimator difficult to live with, however; in defining biserial and point-biserial r in Chapter 15 he reverts to the biased estimator, apparently in an attempt to keep the formulas compact and attractive.

The special strengths of *Statistical Analysis in Psychology and Education* and the features which make it about the best buy on the market are contained in chapters nine through twenty-four. Chapter 9 "Essential Ideas of Sampling" and Chapter 10 "Tests of Significance" of the first edition have become four chapters in the sec-

ond edition: "Sampling," "Estimation," "Tests of Significance: Means," and "Tests of Significance: Other Statistics." Although there is a certain logical elegance in introducing the standard error of $\bar{x}$ for finite populations first and then obtaining $\sigma_{\bar{x}}$ for infinite populations as a limiting case (as Ferguson has chosen to do in Chapter 9), this approach is particularly poor pedagogy. Given that the students have mastered the material in Chapter 7 on the variances of sums and differences of variables, comprehension of a proof that $\sigma_{\bar{x}}^2 = \sigma_x^2/n$ should not be beyond them. Chapter 10 "Estimation" is well done, though somewhat short. Ferguson is to be commended for having discussed the unbiasedness, consistency, efficiency, and sufficiency of estimators. Chapters 11 and 12 "Tests of Significance: Means" and "Tests of Significance: Other Statistics" are fairly complete and correct. The exposition of inferential statistics was improved in this edition; greater emphasis was placed on types of erros, levels of significance, and directional significance tests. Ferguson's treatment of inferential statistics is preoccupied with tests of significance instead of interval estimation. The methods of setting confidence intervals around a statistic are given little emphasis or excluded altogether. Enough has been said elsewhere about the relative merits of these two inferential statistical approaches. Suffice it to say here that such reliance on tests of significance as is evident in *Statistical Analysis in Psychology and Education* is unhealthy.

Chapter 13 "Chi Square" remains unchanged from the first edition. Chapter 14 "Rank Correlation Methods" is actually poorer than it was in the first edition. The section of Kendall's *tau* has been shortened to the point where it is difficult to imagine one learning how to compute *tau* from it. Ferguson added a section on rank correlation when one variable is a dichotomy, but he chose to present a coefficient due to Whitfield instead of Cureton's rank biserial coefficient, which is probably a superior measure. When Chapter 15 "Other Varieties of Correlation" is coupled with the other chapters on rank correlation methods and Pearson's r, few correlational measures of importance are missing from the text. The sixteenth chapter "Transformations: Their Nature and Purpose" is unique. Ferguson has drawn together and unified standard scores, percentile points and ranks, normalizing transformations, the stanine scale, regression transformations, and transformations with age allowances. Though it might be argued that this chapter belongs nearer the front of the text, there is no doubt that it is superbly written and a valuable inclusion.

Chapters 17 through 20 comprise a first-rate primer in experimental design and analysis. The first chapter of this set, "The Structure and Planning of Experiments," is new to the book; indeed it is an innovation in elementary statistics instruction. The

objectives of the chapter are modest (only basic terminology, randomization, etc., are discussed), but it sets the stage well for chapters 18 and 19 in which one-way and two-way analysis of variance procedures are presented. Ferguson's presentation of the analysis of variance is not uniformly good. The statistical assumptions of the technique are handled poorly, being stated imprecisely and incompletely near the end of the discussion of one-factor designs. The independence assumption in the analysis of variance is omitted from the list of assumptions on page 294. This is a pardonable oversight, but Ferguson compounds the omission into a mortal sin in the two-way analysis of variance chapter by analyzing two dependent-groups (repeated measures) designs as though an independent-groups analysis were appropriate. The entire discussion of the analysis of variance could be improved substantially by greater care in treating assumptions. In repeated measures designs, F-tests require spcial modifications; but Ferguson gives no indication of this. On page 273, he points out the virtues of the repeated measures design but never mentions how to deal with its special problems. The effects of violations of assumptions on probabilities of Type I and Type II errors have now been thoroughly charted by mathematical statisticians. The importance of these problems and the success of recent attacks on them are not reflected in the cursory treatment given them on pages 294–295. Ferguson revised a dismal section in the first edition on following the F-test with multiple t-tests; the corresponding section in the second edition is a brief treatment of Scheffe's multiple comparison procedure. The inferential pattern of Scheffe's procedure (error rates are "experiment-wise" rather than "test-wise") is not discussed sufficiently, and Tukey's multiple comparison procedure (superior to Scheffe's for simple contrasts) is conspicuously absent.

Chapter 20 "Analysis of Covariance" and Chapter 21 "Trend Analysis" are valuable additions to the second edition. Very little of the rationale of analysis of covariance is given; but the impact of the chapter on the student is unmistakable: he should be able to *perform* a one-way analysis of covariance and investigate the homogeneity of regression coefficients assumption, as well, after reading the chapter. One can ask for more, but he seldom gets as much. The chapter on trend analysis will be out of reach for many students.

The final three chapters of *Statistical Analysis in Psychology and Education* are essentially unchanged from the first edition. "Selected Nonparametric Tests" does include two sections on nonparametric trend analysis which appear to be original contributions by Ferguson. This chapter is marred only by the absence of sufficiently explicit statements of the statistical hypotheses being tested.

One really has no right to insist that a book be compendious, correct, *and* well written too; but an easy, sometimes humorous style is an added bonus. Ferguson's prose tends to be dry and stolid and, infrequently, annoying (e.g., the repeated use of "comprised of") The gremlins which plague statistics textbook writers left evidence of their work at several points in this second edition. Perhaps the most embarrassing error to the publishers is the misspelling of the name of the consulting editor for statistics (Jonas for Jones). The scatter diagram in Figure 8.1 is inaccurately drawn. The expectation of the F-ratio is not unity given a true null hypothesis, as is claimed on page 319. (The expected value of $F_{n1,\,n2}$ is $n_2/(n_2 - 2)$.) The correlation coefficient is said to equal the geometric mean of b_{yx} and b_{xy}, but the geometric mean is not defined for negative numbers. A question-begging proof that r cannot exceed $+1$ or be less than -1 is given on page 109 (How does one know that r is maximal when the z-scores on x and y are equal for all persons?). We are inclined to disbelieve his complicated expression for the expected value of the mean-square "between" in the one-way analysis of variance (p. 288); the sole expression for $E(MS_{bet.})$ which we can justify, namely $\sigma^2 + \Sigma n_j\alpha_j^2/(J-1)$, equals Ferguson's expression only when all n_j are equal. That Ferguson's formula is equivalent to the expression for $E(MS_{bet.})$ in the random model for unequal n's leads us to believe that he drew an incorrect analogy between the fixed and random analysis of variance models at this point. Regardless of how the error arose, his formula should be disregarded. The statistical *faux pas* are far less frequent in Ferguson's text than in comparable works. The technical accuracy which Binder (1960) commended in the first edition of *Statistical Analysis in Psychology and Education* is evident in the second edition. Ferguson will seldom lead the reader into error.

Binder (1960) concluded that the first edition of *Statistical Analysis in Psychology and Education* was "wholly adequate and, perhaps a superior, text." The revisions and presentations of new material that mark the second edition make the text clearly superior to most competitors. Its continued popularity seems assured.

GENE V GLASS AND THOMAS O. MAGUIRE
University of Illinois

REFERENCES

Binder, Arnold. Review of George A. Ferguson's *Statistical Analysis in Psychology and Educational and Education.* EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 863–869.

Ferguson, George A. *Statistical Analysis in Psychology and Education.* New York: McGraw-Hill Book Co., 1959.

Analysis of Variance by William C. Guenther. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964. Pp. viii–199. $7.95.

This book was written by a mathematical statistician for students with limited mathematics background. It assumes a semester of both pre-calculus college mathematics and statistics. The book is divided into six sections. The first section is a 29-page review of statistical inference. Included in this review is a discussion of *t*-tests, tests for homogeneity of variance, and a number of basic statistical definitions such as Type I and Type II errors and the power of statistical tests. The next four sections of the book are concerned with a number of simple experimental designs such as the simple one-way analysis of variance, random blocks, latin squares, and factorial designs. The sixth and last section of the book discusses analysis of covariance.

An attraction of the book is that many aspects of analysis of variance are discussed which are not usually covered in some of the traditional texts written primarily for psychology and education students. The book itself is not slanted toward any particular audience. The examples are taken from all disciplines. The author spends considerable space in partitioning the total sum of squares for each of the designs. He also discusses the use of orthogonal contrasts for each of the designs and emphasizes the use of multiple comparisons for arriving at the significance level of contrasts. Of particular importance is the author's discussion of the power of the F tests for each of the experimental designs. Most psychological and educational statistics textbooks devote little or no attention to power calculations or to a consideration of the power of the various F tests before the experiment is begun.

The most attractive feature of the book is the differentiation between fixed and random effect models. The author discusses the expected mean squares for both cases and points out the differences in the F ratios between the two models. The concept of interaction is given good coverage. The assumptions for each design are clearly spelled out, and the importance of randomization is continously emphasized. Other useful techniques such as tests for nonadditivity and efficiency calculations are also covered briefly.

The major shortcoming of the book is that it is written rather concisely with brief discussions of a large number of concepts. Some of the formulas look rather formidable and this might discourage some students. The book by itself would probably not suffice as a text in experimental design for psychology and education students because of its consciseness and limited coverage. However, it would be an excellent supplemental text or reference book to use in emphasizing some of the points to which the more traditional texts give little or no treatment.

George H. Dunteman
University of Florida, Gainesville

Nonparametric Trend Analysis (A Practiced Guide for Research Workers) by G. A. Ferguson. Montreal: McGill University Press, 1965. Pp. 61. $1.95, paper.

In a fascinating little monograph—albeit a rather expensive one—Ferguson has created a guidepost by which the practical researcher may lead himself through the paths of nonparametric analysis for trend.

Ferguson needs no introduction as a lucid, careful writer. In the present booklet he does nothing to destroy that image—in fact, this is Ferguson at his clearest. For the most part, the printers have done an excellent job for him. The shape of the monograph is a trifle awkward perhaps, but the greatest drawback in format is the inclusion of tabular reference material in the body of the text. Having to hunt for a table is always time-consuming, at best. One final comment on the format—the binding would appear to be inappropriate to the kind of "recipe-book" function the monograph sets itself. The review copy was—after two mailings—already in a state of dismemberment. Presumably the book would not last the user who attempted to open it fully, much longer. It is a pity that a book with such a heavy reference-orientation cannot be bound to make it easier for desk use.

Nonparametric trend analysis parallels, for ordinal data, the parametric technique of orthogonal polynomials. That is, it is a method designed to facilitate the study of the relation between treatment and experimental variables to see whether it is monotonically increasing or decreasing, and what the general shape of the functional relationship is. As with other nonparametric techniques, the methods described in the monograph void themselves of the usual assumptions of the analysis of variance model which they replace—i.e., no assumptions of normality or of homogeneity are made. Because of the nature of the data on which it relies, nonparametric trend analysis is potentially highly appropriate for the social sciences.

The Ferguson monograph presents a complete method for conducting the analysis. Basically, it is an extended presentation of an application for the sampling distribution of the statistic S, as used in the definition of Kendall's rank correlation coefficient, *tau*. The reader will recall that S is descriptive of lack of order in a set of paired ranks. If one set of ranks is arranged in order, the degree of disarray in the other set is given by S, which is a summed set of weights over the $n(n-1)/2$ possible comparisons for n ranks. If one pair in the comparison is arranged in natural order, a weight of $+1$ is assigned; if not, a -1. These ± 1's are summed to give S (any standard text sets out the procedures for ties). A positive value of S means that the ranks of the second set tend to increase monotonically with an increase in the naturally-ordered set; a negative value in-

dicates that the ranks of the second set show a tendency to decrease monotonically. S is a particularly appropriate statistic—its asymptotic normality and its power, were investigated by Mann twenty-one years ago. Problems involving ties have, in general, been fully explored. And one of the most powerful nonparametric tests, the Mann-Whitney-Wilcoxon U is yet another application of S.

Chapters 1–3 in the monograph (covering the first 18 pages, or 30 per cent of the space) are introductory. Chapter 1, in fact, describes how the monograph was written in response to some problems encountered in making an analysis of data from experiments conducted in the Department of Psychology at McGill University. Chapter 2 defines the terms used, and Chapter 3 summarizes the computational procedures for S and for the variance of its distribution. In Chapter 3, the author caters to the researcher by ending the section, and footnoting each table, with sets of rules. Some of the phraseology in these rules is perhaps rather terse for a nonstatistician: e.g., "when one variable is a dichotomy and the other contains m groupings of values of extent t_i where $t_i \geqq 1$, reduce the absolute value of S by the correction . . ." (p. 18). One can imagine an unsophisticated researcher looking for a set of simple directions, being unwilling to read through the text, and being more than a little bewildered at the technicality of such language.

Chapters 4–6 set out the tests for monotonic trend with correlated data. Chapter 4 is concerned with large-sample procedure without ties. The suggested method is an alternative to the Friedman two-way analysis of variance, by replacing the original measurements with ranks. Chapter 5 treats the case of large samples with ties; Chapter 6, small samples. Chapters 4 and 5 contain excellent examples laid out as tables, but which merit only rather disembowelled mention in the text. It could be helpful to know what the data are—to give the researcher some notion of the kinds of variables for which the analysis has been found helpful.

Chapter 7 sets out the procedures for testing for monotonic trend in independent samples. S is calculated in the usual way from the data set out in a manner similar to a Kruskal-Wallis one-way analysis of variance, and using a normal score transform. By now, the reader who is familiar with any compendium of nonparametric tests, will be appreciating the unification brought about by use of S. It is not always a simplification, but the interredatedness of many familiar tests is brought out very clearly.

Chapter 8 returns to a theoretic discussion of orthogonal polynomials and the concept of nonmonotonic rank correlation. In the latter, ranks of a set Y are compared with the natural ordering of ranks on a set X. Monotonic rank correlation is, then, a measure of the disarray of Y in relation to the ranks for a first-degree orthogonal polynomial. The possibility of extension to second-degree

and higher-order polynomials is considered. An extensive Chapter 9 works through illustrative examples in the case of correlated data; Chapter 10, for independent samples.

Here, then, is a satisfying monograph bringing together under one statistic a number of techniques already available for the most part, but perhaps not so well set out for the practical researcher. It is to this audience that one commends the book. And certainly it deserves a corner in the bookshelf of statisticians in the areas of psychology and education.

PETER A. TAYLOR
Rutgers University

Readings in Educational and Psychological Measurement by Clinton I. Chase and H. Glenn Ludlow (Editors). Boston: Houghton Mifflin Co., 1966. Pp. xi + 381. $4.25

Instructors of courses in measurement and evaluation are fortunate in having a number and variety of rather good texts on which to draw. However, instructors do have a problem in finding ways of supplementing the basic text. One can assign students to read articles in the various journals, but this creates the scene of students falling over one another to obtain the library's one copy of the particular journal. The usual upshot is that the journal mysteriously disappears or someone discovers that the article was neatly removed with a razor blade. A second alternative is to prepare mimeographed copies of supplemental readings. This is fine from the student's point of view, but requires a tremendous amount of secretarial time—a commodity in rather short supply in most places.

Chase and Ludlow, responding to a real need, have prepared a paperback book of supplemental readings. The book consists of forty-four articles, research reports, and "pieces," which should go a long way towards satisfying most instructors and, hopefully, students. It is a delicious collection, and most instructors will be glad to have so many important readings gathered into a single volume. The selection has been thoughtful, and the editors wisely limit their introduction to one or two short paragraphs, so as to allow maximum opportunity for the authors to speak.

The articles are grouped into seven units, but this arrangement is relatively unimportant. It is the articles themselves that are important, and, fortunately, there is little to interfere with them. Perhaps the best way to convey the usefulness of this volume is to present a partial list of the contents. These include: Tyler's "General Statement on Evaluation," Cronbach and Meehl's "Construct Validity in Psychological Tests," Mosier's "Problems and Designs of Cross-Validation," Honzik, Macfarlane, and Allen's "The Stability of Mental Test Performance Between Two and Eighteen

Years," Coleman and Cureton's "Intelligence and Achievement: The 'Jangle Fallacy' Again," McNemar's "Lost: Our Intelligence. Why?," Wesman's "Faking Personality Test Scores in a Simulated Employment Situation," and Super's "The Multifactor Tests: Summing Up." There are many other equally fine articles.

The only objection this reviewer would raise, albeit a minor one, is the lack of historical material. For example, inclusion of some of Binet's, Ben Wood's and E. L. Thorndike's writings could have added an important historical perspective for the student. However, treat this objection lightly. The collection is first rate and deserves thoughtful consideration by every instructor of courses in evaluation and measurement.

RICHARD WOLF
University of Southern California

A Practical Introduction to Measurement and Evaluation (Second Edition) by H. H. Remmers, N. L. Gage, and J. Francis Rummel. New York, N. Y.: Harper and Row, 1965. Pp. xvii +390.

From their acquaintance with previous works of the authors, the reviewers had higher expectations for this book than were realized. The second edition differs little from the earlier version; brief sections on item analysis, the affective domain, and the standard error of measurement constitute the major changes.

Although billed as a "practical introduction," the book evidences no particular strength in this respect. It appears to be overly ambitious in scope, touching topics which range from measurement theory through mental hygiene. As a result, several important concepts are either minimized or over-simplified.

Too frequently the reader is given an admonition without concrete assistance as to how to carry it out. Note the following illustrations:

> Pupil evaluation should be used as a basis for improving public relations and the mobilization of public opinion (pp. 11–12);
>
> . . . the teacher's knowledge of the pupil should guide the decision of the pupil and his parents regarding the type of secondary education to choose. (p. 91);
>
> . . . the emotional and social adjustment of pupils should be informally evaluated frequently. (p. 108).

A more severe limitation of the volume results from the lack of careful updating of content. A number of important omissions or oversights are indicated by these examples: The Wechsler-Bellevue is mentioned but not the WAIS; the Alice and Jerry Readiness Test and the Kuhlmann-Finch Test appear, yet neither the ITBS, Stanford, Metropolitan, nor California Achievement Series is men-

tioned in the list of typical achievement tests; NCME is presented as the former NCMUE; the *Journal of Educational Measurement* is not included along with six other journals listed as carrying reports of research with tests; some of the suggested references appearing at the end of each chapter are not the most recent editions; and finally, the obsolete probable error is defined and mentioned several times throughout the book. These examples, along with the general tone of the presentation, serve to indicate a lack of care in updating in content and emphasis.

Chapter 2, which presents an introductory survey, unfortunately becomes an extended glossary of types of tests rather than a general orientation to measurement and evaluation. This feature may have an initial negative impact upon the reader. The chapter, "Statistical Concepts Used in Measurement," although comprehensive, is not organized in a highly teachable manner. The care taken in presenting steps for grouping data to facilitate the computation of the standard deviation is somewhat inconsistent with the raw-score approach used for determining a coefficient of correlation. Too much information of a tangential relevance may hinder the students' acquisition of the major concepts; for example, three different methods for determining split-half reliability are given. Certain points are either imprecisely or incompletely made. For example: T-scores and stanines are not presented as being normalized; ordinate is used where abscissa is implied (p. 56); no mention is made of the inappropriateness of using the Spearman-Brown formula with speeded tests. Two scatter plots illustrating correlation are improperly scaled; the range on one axis is about four times as great as on the other. There is no mention of the widely used deviation IQ. A more serious error appears in the definition of a "true score" which is said to be ". . . a score that would be obtained by an ideal, infallibly accurate measurement." (p. 134). The definition obscures the very important point that constant error is an ingredient of a "true score." Further, there is an unfortunate promulgation of the widespread misconception that a test, in order to be maximally efficient, needs a wide range of item difficulty (p. 203).

The sections on validity and reliability are generally good. However, a few issues are apt to lead the beginner astray: (a) "face" validity is presented as a synonym for content validity; (b) the fact that reliability can be increased at the expense of validity is overdone to the extent that the student may perseverate on this concept rather than on the more typical pattern, namely, that they are generally correlated; (c) the multiple-choice test is stated as being free of response sets, ignoring its vulnerability to the potent "gambling" and "speed vs. accuracy" response styles; (d) the diagnostic value of profiles is both implicitly negated (p. 112) and supported (p. 113); and (e) classroom teachers are told that they

can standardize their tests by applying statistics (p. 95). Unfortunately, little attention is given to the meaning and interpretation of reliability and validity data.

The chapter on Test Administration and Scoring is very good, although a clearer distinction between suggested teacher-behaviors for standardized versus teacher-made testing could profitably have been made.

The main strength of the book is found in Part Three: The Evaluation of Classroom Instruction, in which a good discussion of evaluation in relation to necessary consideration of educational objectives is presented. This section, unlike the others, is supported by relevant documentation and useful illustrations. The relative merits of different approaches to testing, for example, objective versus essay, provide practical guidelines for improvement of teacher-made tests. It provides a good source of examples for construction of various types of test items. The attempt to deal with the complexities of assigning marks and reporting pupil progress adds little new insight to an everpresent problem.

Part Four: Appraisal of Personality Aspects is a mixture of mental hygiene vocabulary and concepts and an inadequate and uncritical treatment of interest and personality assessment. Curiously, the Vocational Interest Inventory and the Occupational Interest Inventory are given as much emphasis as the Kuder and the Strong. The Allport-Vernon Study of Values is presented as interest inventory. It is unfortunate that the authors recommend adjustment inventories for general pupil use. Thus, they give inadequate attention to the problem of validity and level of training which is necessary to properly utilize such instruments.

The reviewers regret that the attention, care, and quality which the authors have manifested elsewhere in their work was not reflected in the present effort.

KENNETH D. HOPKINS AND WILLIAM A. SEASE
University of Colorado

Building a School Testing Program by Robert H. Bauernfeind. Boston, Mass.: Houghton Mifflin Company, 1963. Pp. xvii + 343. $5.25.

This text is designed specifically for school people who are responsible for deciding which tests to use at which grade levels. Thus, the book is intended primarily for school administrators, testing directors, and for graduate students in programs that may lead to their becoming school testing directors. Although the author assumes that the reader has a basic knowledge of educational measurements, he takes the position that statistical methods are not crucial to the work of school testing directors. The result is a text which recognizes the importance of the statistical concepts

inherent in measurement, but treats them conceptually rather than computationally.

The text is organized in three major parts. Part I emphasizes the basic concepts of test scores, test norms, correlation, validity, and reliability. Although the overall discussion of the concepts lacks depth, Bauernfeind does an exceptionally able job of providing the essentials needed for use in test selection and implementation. In regard to establishing local norms, this is one of the few texts which advises strongly for the use of stanines over other methods of converted scores. The discussion of the advantages and disadvantages of using stanines which is done quite well should be understandable even for an unsophisticated school testing director.

Part II is clearly the strongest aspect of the text and perhaps the most useable for those who are responsible for initiating a testing program. There are chapters on testing in grades kindergarten through three, testing for educational development and for mental ability in grades three through twelve, testing for vocational aptitudes and vocational interests in grades eight through twelve, measuring personality characteristics, and using subject-matter achievement tests. In this part especially, Bauernfeind does not hesitate to offer specific advice and guides to action. As a whole, the suggestions are sound, and a school testing director would be unlikely to go wrong if he were to follow them without question. However, Bauernfeind's willingness to prescribe may occasionally place a competent course instructor in an awkward situation, since for various reasons the instructor may at times prefer alternate or different emphases from those prescribed in the text. However, the flavor of the volume makes this event unlikely. A potential user of the text would, of course, have this problem resolved prior to requiring the book for course use. As is the risk in using any text, if the instructor raises too many questions, he may undermine the students' confidence in the text, himself, or testing in general. Thus, for the most part, if a book is used, he will have to join with Bauernfeind in most of the suggestions given.

A concluding chapter illustrates an example of how one school's testing program was revised through using the recommendations and procedures emphasized in the preceding chapters. Actually, the material which preceded this chapter provides all the cues and prompts necessary for establishing an effective testing program and tends to make the chapter redundant. A greater detriment may be that the less motivated school testing director will go directly to the chapter and cook an "instant" program without using the ingredients suggested in the preceding chapters to generate his own testing program recipe.

A unique inclusion in the text is a section in which nine leading test specialists attempt to forecast future developments in educa-

tional testing. Although the forecasts are provocative, they add little to the immediate objectives of the book. However, they do tend to illuminate trends and issues to be concerned about in the overall domain of educational testing.

Noticeably lacking in the text is any functional information on how data processing techniques might be implemented in building a school testing program. Unfortunately, the author dwells only passively with the subject and even then in an admittedly suspicious fashion. This gap is even more pronounced when one reads of the importance attached to data processing by each of the nine contributing test specialists.

Overall, the text is functional, and well presented. As Bauernfeind acknowledges, some measurement specialists may have reservations about the nonstatistical nature of the book and about some problems and concepts that have been unduly simplified. However, there is little evidence that highly sophisticated texts attempting to do the same thing have had any great observable impact on "building a school testing program." Thus, it is the impression of this reviewer that the shortcomings cited are well outnumbered by the strong points.

James C. Moore
The University of New Mexico

Individual Differences by Anne Anastasi (Editor). New York: John Wiley & Sons, Inc., 1965. Pp xiv + 301.

This excellent little volume affords the measurement expert the opportunity to sample nearly a century of the most pertinent publications in the field, from Galton to Guilford. The thirty titles are interspersed with editorial passages which do just what Anastasi hoped for. They act "as a guide who takes the reader on a personally conducted tour of the literature."

Individual Differences will prove a valuable book for teachers, because it provides materials which are not readily available in many libraries (e.g., three of the Binet and Simon articles and several Galton works). This reviewer observed that almost all the works have been listed by Anastasi as references in her earlier differential psychology text; she must have made her choices of these particular entries partly because they have enabled her during her years of teaching to best put "the psychological study of individual differences in historical perspective."

The first chapter is a succinct survey of that which went before and led up to the scientific study of individual differences. Each of the following chapters contains from two to four chronologically arranged and annotated readings on ten different topics, concerning or closely related to measurement. The ten topics are: Measurement of Individual Differences, Nature of Intelligence: Pioneer

Research, Nature of Intelligence: Later Developments, Behavior Genetics: Statistical Studies, Behavior Genetics: Theoretical Orientation, Cultural Deprivation: Cross-Sectional Approaches, Cultural Deprivation: Longitudinal Approaches, Nature of Genius: Early Exploration, and Nature of Genius: Focus on Creativity.

The concluding chapter briefly gives an overview of the selections as they relate to the story of the progress and expansion of knowledge about individual differences. The editor points out that some of the selections were chosen to demonstrate that false starts have been made and that the "history of differential psychology has been particularly productive of controversy."

All in all, students and teachers of measurement as well as researchers with an appreciation of history's impact on the present should find this a valuable volume to own. The selections are well chosen. They are not so highly technical that they cannot be read with ease; they are interesting; and the editor's comments tie the selections together so as to lend a certain charm often found missing in scientific writing.

Sister M. Jacinta Mann
Seton Hill College

The Psychology of Individual Differences (Third Edition) by Leona Tyler. New York: Appleton-Century-Crofts, Inc., 1965. Pp. x + 572.

Once again Leona Tyler has made a theoretical and practical contribution to the area of individual differences. In her usual clear, interesting, and informative style, she has managed to synthesize the overwhelming amount of material in the area into an extemely readable and carefully organized textbook.

Nine years have elapsed since the second edition of her book. The format of the third edition is similar to that of the first two, but much updated. Again, as in the second edition, there is a lesser emphasis on the presentation of statistical concepts than there had been in the original textbook. The concepts presented form an integral part of the text. They are not an added appendage. The logical fashion in which Part One proceeds from historical considerations to general principles to research strategies provides a good prelude to the specific research findings presented in the later chapters.

The separate presentation of individual and group differences makes for greater clarity. Presented in any other form, such numerous findings could indeed be confusing to the student. As it stands, the material is not only logically presented, but also provocative of interest in the reader.

This reviewer notes that any personal bias that the author might have with regard to the findings is held in abeyance. Two chapters seem particularly impressing and pertinent: Chapter 9

on cognitive styles and Chapter 12 on race differences. The area of cognitive styles is very much to the fore at the present moment. This chapter gives a much-needed summary of the cognitive research to this point. The author sees the research of Witkin, Holtzman, Klein and many others as occupying a "strategic position in differential psychology."

The chapter on race differences is especially good. The author neither wholeheartedly accepts the notion that one race is inferior to another, nor does she deny the empirical existence of such differences. She simply calls for research concerning the possible causes. Throughout the chapter, there is developed a real rationale for the organization of poverty programs, especially for the programs which will reach the preschool culturally deprived youngsters. As she states in the Preface to the book, much of the material in the textbook should be related to the social reforms of our time.

After treating individual and group differences, the author discusses the factors which produce these differences. She dispels, once and for all, the idea of heredity *or* environment as the sole contributing factor. While treating each of these factors in separate chapters, she nevertheless maintains the necessary interaction between the two.

The final chapter "The Science of Human Individuality: Past and Future" gives a brief survey of some of the concepts which have arisen from the cited research, including the concepts of *strategies* and moderator variables. Appropriately enough, the chapter ends with a blueprint for the future.

The Psychology of Individual Differences should prove helpful to the instructor, and enlightening to the student or proponent of differential psychology. This reviewer found it to be just this.

Sister Mary Walter Hauk
Seton Hill College

Two-Person Game Theory: The Essential Ideas by Anatol Rapoport. Ann Arbor: The University of Michigan Press, Ann Arbor Science Paperbacks, 1966. Pp. 230. $1.95.

As a technique that has considerable potential usefulness in the social sciences, decision-or game-theory has received relatively scant attention. Recent appearances—in increasing volume—of articles illustrating its applications, may hopefully imply that game theory is at last "catching on" amongst psychologists and educationists. Within these two fields alone, the list of persons who could benefit from an acquaintance with game theory is too extensive to begin—anyone who has to make a decision in a conflict situation would warrant inclusion. It is pleasing to note, therefore, that this new little book, simple and readable, is available.

Rapoport himself is a mathematical biologist. In preparing his

text, however, he has kept to a bare minimum any mathematical skills required to appreciate and to understand the content of the book. It is to this clarity and simplicity of expression that the book owes its values as a contribution to a widening of the social scientist's horizons. For the most part, high school algebra, and a very meager knowledge of analytical geometry would more than suffice the reader. To be able to reduce what is a complex, and weighty field to this level of difficulty is in itself an achievement; and moreover to maintain interesting dialogue is an indicant of literary merit.

The restriction on the mathematical sophistication demanded of the reader necessarily also restricts the scope of the book. The author has found it possible within the confines he has set to present the essential ideas of game-theory in the context of the two-person game. He leaves the general case for a later volume. By making this choice, notational simplicity is maximized, and at the same time important topics may be treated in some depth.

By a neat analogy, game theory is defined as being to games of strategy what probability theory is to games of chance. Game theory then, is not in the position to answer questions such as, "What is the best way to play bridge?" but accepts questions of the kind: "Is there a best way to play bridge?" Game theory is, accordingly, primarily concerned with the classification of games—and in this, has much in common with other classificatory sciences.

Here, then, was a golden opportunity for the author to present to his audience—allegedly social scientists—examples and applications in the fields of sociology, psychology, and education, for instance. Yet, few direct applications appear. Almost all the illustrative examples in the text are confined to familiar games (such as tic-tac-toe), and to the eager social scientist, this lack of demonstration that game theory *can* be appropriately used in his field, may be something of a frustration. A 30-page chapter devoted to "opportunities and limitations" does little to relieve that feeling. Surely, few amongst us can find solace in the assurance that two person game theory is highly applicable to war situations.

On the positive side, however, one can find much to commend. The first chapters of the book are extremely well written and present—at their intended level—the essential characteristics of games as distinct from decision theory (Chapter 1); the basic notions of utility (Chapter 2) and strategy (Chapters 3–6). Even to the most ardent mathematical-symbolist, there is a delight in seeing familiar notions reduced to words with such skill. The chapter on utilities is a particularly appealing one; the sections on scales of measurement and payoffs as utilities are especially so. But it is not until Chapters 3–6 that there will be much that is new to the "average" social scientist. The ideas of simple and of mixed strategy are introduced (Chapters 3, 6) by reference to "tic-tac-

toe" and "square the diagonal" respectively. Although these chapters are clear and quite devoid of any but the most elementary mathematics, they become a little tedious—but this is probably the inevitable outcome of replacing mathematical symbols with words. Chapter 4 sets out the notions of the game tree and the game matrix; Chapter 5 discusses "dominating" strategies and the concept of minimax. A dominating strategy, s_d, is, of course, one that will do at least as well as any other strategy, s_k, no matter what strategy an opponent adopts. Minimax—relatively easy to explain mathematically—looks a little awesome in verbal garb, but the author handles the topic well.

Chapter 7 discusses the two-person zero-sum game. In effect, the underlying theory here tells us that one is not attempting the impossible in looking for a best strategy for each player. The algebra is a little heavier in this chapter, but still about grade 10 level. Chapters 8 and 9 are particularly interesting. They discuss negotiated and non-negotiable games. The former of these (which arises when both parties hope to gain from the results) would seem to have particular application to, say, educational administration. The latter depends on a theoretical development which is quite similar to that of stochastic learning theories.

Chapters 10 and 11 develop dynamic game models and in the process call on a smattering of differential calculus and analytic geometry. In a sense, these are the most enlightening chapters because of their greater generality.

Here, then, is a little book that many social scientists should be encouraged to read. It can scarcely be regarded as a text—only the most fortunate of universities at present have courses in education and psychology the content of which approaches that of this book. It is, then directed at an interested, semi-specialist audience. In providing the basic notions of two-person game theory, the author would seem to have achieved his objective admirably, and with remarkable clarity of expression. The book will not long appeal to those who are mathematically inclined—it is a litlte tedious for that. But then, it is not intended for the mathematician; and if it can whet the appetite of the administrator, the counselor, the clinician and the like, to seek further for possible applications of the method, the book will have made a major contribution. But, if those same people are looking for examples in this volume that are allied to their own field, they will be disappointed.

PETER A. TAYLOR
Rutgers University

Pattern Recognition: Theory, Experiment, Computer Simulations, and Dynamic Models of Form Perception and Discovery by Leonard Uhr (Editor). New York: John Wiley & Sons, Inc. 1966. Pp. xii + 393.

Pattern recognition has been an unpopular area of psychological inquiry, but recent results from neurophysiology and from computer science have given new impetus to the area. Uhr has collected an interesting set of twenty-two readings that sample these new results as well as the philosophical, empirical, and theoretical background of the problem.

The selections have been sorted into five parts. Part I concerns the conceptual framework of the problem and contains contributions of a philosophical nature by C. S. Peirce, E. Cassiere, K. Craik, and L. Wittgenstein. These selections contain material that is thought-provoking and fascinating, but taken as a whole they are fragmentary and do not provide a coherent, complete account of the nature of pattern perception. Perhaps that is as it should be in light of the current state of the problem. Part II consists of three excellent, comprehensive reviews of different aspects of the psychological literature. M. D. Vernon considers the literature in terms of the fundamental stages in the process of perceiving; J. F. Wohlwill reviews development studies of perception, and H. W. Hake reviews empirical work on form perception. The Hake selection is especially welcome, since it was previously available only in a Wright Air Development technical report. Also in Part II is the important procedural article by F. Attneave and M. D. Arnoult, describing their extensively used method for the random generation of patterns. Part III includes theoretical approaches to the problem of pattern recognition by J. A. Deutsch, P. C. Dodwell, W. H. Marshall and S. A. Talbot, and W. Reichardt. The latter two papers are especially interesting in showing how a good theory can integrate a large number of empirical findings. Part IV contains four papers discussing some findings from the neurophysiological investigation of vision in frog, cat, and octopus. These papers, by H. B. Barlow; B. D. Burns, W. Heron and R. M. Pritchard; D. H. Hubel and T. N. Wiesel; and J. Z. Young, are all concerned in one way or another with the processing, transmission, or storage of patterned information in the nervous system. Part V contains five papers describing computer algorithms for the automatic recognition of patterns and one general paper by Uhr discussing the general nature of automatic pattern recognition. The paper by L. G. Roberts provides an interesting insight into the much advertised Perceptron. Three other particular recognition systems are described by W. W. Bledsoe and I. Browning, by R. L. Grimsdale, F. H. Sumner, C. J. Tunis and T. Kilburn, and by Uhr and Vossler, while O. G. Selfridege describes a general framework for such pattern recognition programs. Viewed from historical perspective, it is especially interesting that the computer experts chose to work with upper case letters of the alphabet, a set of stimuli about which very little is known psychologically. The behavioral and model building approaches thus have not made contact as yet.

Nor have the models described here made much contact with the neurophysiological findings of Part IV. Similarly, the simple contours that are generally used in the newer neurophysiological work are not much used in behavioral studies. Perhaps this interesting juxtaposition of the behavioral, neurophysiological, and computer simulation approaches to perception will encourage a closer relationship in the future.

The selections reflect, of course, the editor's own point of view, which is that of a formal theorist with a preference for information processing models. Most of the selections show an interest in what information is transmitted by the sensory system and in how the system processes such information. But the selections are relevant and interesting from any point of view.

Uhr remarks in the preface that pressure to keep the size of the book within some reasonalbe limit forced him to exclude many fine papers. Apparently, he also felt compelled to keep his editorial remarks very brief. They are confined to only one or two pages at the beginning of each of the five parts. Otherwise, the papers are left to speak for themselves. Unfortunately, they do not speak in a coordinated way—the editor should have provided more setting for each paper. Also missing from the collection is any identification of the authors. Particularly in a multidisciplinary volume including philosophers, psychologists, neurophysiologists, computer scientists, engineers, and mathematicians, some background information on the participants would have been very helpful. The reviewer's final complaint is that in his brief commentary Uhr frequently compares or contrasts one of the included selections with some other work that was not included, the latter being identified only by some such phrase as "Steven's work on specifying and programming Deutsch," and there is no further reference anywhere in the book. Such name dropping may have its uses in face-to-face verbal interchange, but in a book it is an annoying discourtesy to the reader.

Psychologists who are developing information-processing models of pattern recognition will find the book very useful. Interested psychologists will find it fascinating. Teachers may have difficulty using the collection in courses because considerable interpretation will be required to fit the articles into a coordinated framework. Nevertheless, by judicious selection many teachers will be able to use a significant set of the readings in a course in perception.

Bert F. Green, Jr.
Center for Advanced Study in the Behavioral Sciences

Computers and the Human Mind; An Introduction to Artificial Intelligence by Donald G. Fink. Garden City, N. Y.: Doubleday and Co., Inc., 1966. Pp. vii + 301.

This book, one of the Science Study Series published by Anchor

Books, has two major objectives. The first is an introduction into the world of computers, and the second is an introduction into the field of artificial intelligence. It is not clear to this reviewer which objective is being sought by various parts of the volume. The Science Study Series is designed for the lay public (as indicated on page vii of this book), and therefore this volume must be analyzed according to its success in presenting, explaining, and perhaps convincing a lay (although intelligent) reader about computers and artificial intelligence. If a reader approaches *Computers and the Human Mind* with no prior experience on computers, and with perhaps limited mathematical background, he will be unable, and unwilling to find his way through the explanations offered. Although the book begins in a very elementary style, it soon presents examples far more complicated. It would seem unlikely that a lay reader would be educated in computers in the first 17 pages sufficient enough to understand pattern recognition, and the immediate comparison of mind to machine (the reviewer would prefer the word "brain") in the first chapter is likely to alienate a considerable proportion of the population at the outset.

There are a considerable number of lapses throughout the volume into language and mathematics for which the reader is offered no explanation. Examples of the derivation of π to 10,000 places, are not presented in a manner which would convince even the initiated as to why one would want to do such things. This book, therefore, is *not* applicable as a means for explaining the functions and uses of computers to the public.

With respect to its second criterion, that of artificial intelligence, the author seems to be more on his own ground. The presentation of human physiology and thought processes is quite well done for the type of audience likely to read it. The analogies to computer "thought" are less convincing, but those in the part on translation are well presented.

The book suffers from too many styles, from a rather uninteresting presentation of an intriguing and challenging subject, and from poor organization. The contents could have been "programmed" more adequately. The volume presents too much information too quickly, and tries to serve too many masters. It is too elementary for a text book, and too advanced for an average reader.

The author should be commended, however, on the mass of data and information presented in the 300 pages. The information is accurate, and the chapters on the physics of computers are well written. The book can be recommended for "outside reading" in a basic psychology course, and certain chapters are useful for the specific presentations therein.

RICHARD E. SPENCER
University of Illinois

The Tools of Social Science by John Madge. New York: Doubleday Anchor Books, 1965. Pp. ix + 362. $1.25.

As Deputy Director of Political and Economic Planning in England, John Madge directs sociological and economic research for government programs. It is important for the reader to know that the author's forte is sociology; the book's title, however, demonstrates that the author adopts an interdisciplinary approach. Thus, his belief is supported that there are certain research tools common to all of the social sciences. The main social sciences cited in the book are sociology, anthropology, political science, and economics. The author also considers history to be a social science.

The specific social science tools discussed in separate chapters are documents, observation, the interview, and experiment. Although the author feels that all these tools are valuable, it is clear that they are ordered in such a way that experimentation is considered to be most desirable, i.e., the most distinguishing feature of a mature science. The problems connected with each technical research tool are discussed by the author. The first chapter of the book deals with the method of social science; special attention is given to the language and the logic employed in the social sciences.

Since Sputnik I, institutions involved in teacher training have given a great deal of attention to the research methodologies of the various social science disciplines and of history. This emphasis has been especially true in the area of social studies education. How one knows what one knows is often considered as important or perhaps more important than what one knows. There is, therefore, a need for books which focus on this subject. Although there are many criticisms which may be made against John Madge's book, the book may prove to be of heuristic value to the reader interested in social science research methods.

The most pointed criticism of the book is that it is dated: the most recent source cited is fifteen years old. The paperback edition has not been updated but is in the original 1953 form. Yet it is precisely during the last fifteen years that there has been a surge of activity in questioning previously used social science research methods and in forging ahead to develop new methods.

It is also highly questionable that a sociologist can speak for social scientists in other disciplines as to research methods. This becomes increasingly true as each discipline becomes more sophisticated in developing methods of inquiry. Historians would probably be annoyed at the author's inclusion of their discipline in the social sciences. In short, the interdisciplinary approach adopted by the author should be carefully scrutinized by the reader.

Finally, let it be said that the author may be labelled fittingly as a "traditionalist" in his approach to social science. That is, he does not support the "positivists", or as he calls them, the naive be-

haviorists, who hold that all science is value-free. The reader may find the discussion of this subject as it appears in the book's Introduction to be of more than passing interest.

Dale L. Brubaker
University of California
Santa Barbara

The Rand McNally Handbook of Education by Arthur W. Foshay (Editor). Chicago: Rand McNally and Company, 1963. Pp 294.

As implied by the title, this is a reference-type publication offering many principal facts about education in the United States. The major areas treated include (a) organization and administration, and (b) curriculum at the local, state, and national levels. In addition, descriptions of the educational systems of England, France, and Russia are presented for the reader who has an interest in comparative education.

The volume provides huge quantities of factual information offered in an unbiased, well-organized, and succinct manner. It should provide a base from which anyone who has the interest could develop a fairly comprehensive overview of school administration, organization, and curriculum trends in the United States.

It is quite likely that this work will prove of considerable value to legislators, school board members, school administrators, teachers, interested citizens, and students of education, as well as to a wide range of individuals and groups from other countries. It would indeed make an excellent desk reference.

Obviously, there are a number of thorny problems inherent in deciding upon scope, limitations, and organization. Oversimplification is a possible pitfall in an undertaking of this type. However, the editors have made a noteworthy effort to recommend additional readings for the person who is attempting to develop a thorough understanding.

Another limitation of a work of this type is the fact that many of the statistics shown are outdated even before publication. At least a continual re-editing effort will be necessary to minimize obsolescence by updating statistical data and reanalyzing changing curriculum trends.

The book is divided into three major parts with contributing editors, specialists in each area, assuming the major writing responsibility.

Part I, edited by Arnid J. Bucke, Director of Studies, New York State Teacher Association, is devoted to organization and administration of education. In Chapter I, some of the topics discussed are the role of the federal government in education, a legal basis for education, a description of the U. S. Office of Education and of other federal activities in education. Numerous national statistics about personnel, school plant facilities enrollments and expenditures are in-

cluded. Brief but effective statements regarding current issues with arguments *pro* and *con*, and additional references for further study are given. The editors have encouraged the reader to "whet his appetite and to go on from there."

Chapter 2, much of which is in tabular form, is concerned with a lengthy and technical description of all fifty states. Officials from each state were asked to check for correctness sections dealing with their respective states. Attention is given to a number of areas, ranging from constitutional provisions and legislative control to teacher certification and current issues in state school administration.

A discussion of local school systems with supporting data is presented in Chapter 3.

The major portion of Part II, Chapters 4 through 15, deals with the major subject matter areas which have traditionally constituted the school curriculum. One chapter presents a brief discussion of curriculum methods and organizations in addition to some illustrative daily school schedules. Attention is also given to current curriculum trends and issues.

Chapters 17, 18, 19, 20, and 21 are concerned with audiovisual aids, textbooks, school libraries, adult education, guidance, and vocational education.

In Chapter 22, several digests of controversial statements on educational practices which offer promise for improvement are presented. Concepts such as team teaching, independent study and ability grouping are given attention.

According to the editor of Part III, Frank G. Jennings, editor at large of *The Saturday Review*, the section on education in England, France, and Russia is offered because of "greatly increased interest in the U. S. during recent years." The contributing editor points out the difficulty in making any comparison of education in the United States with foreign countries and stresses that each educational system may best be presented in its own frame of reference.

Frank C. Emmerling
Comprehensive School Improvement Project
Department of Public Instruction
Raleigh, North Carolina

Group Values through Children's Drawings by Wayne Dennis. New York: John Wiley and Sons, Inc., 1966. Pp. xiii + 211.

Readers of EPM are familiar with assessment procedures employing drawings of human figures: Machover's Draw-A-Person (DAP) and its predecessor, Goodenough's Draw-A-Man (DAM) with a modern revision by Harris, and others. Dennis offers a third approach and bases its principles and purposes upon neither of the earlier drawing techniques. While Machover claimed to measure "global personality" and Goodenough measured intelligence, Dennis

stakes out a claim for measuring not the values of the drawer, but the values of the society to which the drawer belongs. Most readers will agree that the author's claim which deserves careful consideration should not be lightly dismissed.

In essence, Dennis makes only two assumptions in the application of his procedure. The first fits in well with projective theory: the drawn figure is derived from internal factors (values) rather than from figures observed in the environment. Considerable attention is paid to this point and the evidence, a mixture of observation and experimentation appears to be ample. The second assumption contradicts the projective hypothesis. Dennis assumes that *favorable* attitudes are indicated for the figure drawn. Negative attitudes toward certain men (Dennis, like Goodenough, analyzes the drawing of the male) are indicated by one of two methods: (a) the absence of drawings of such men or (b) presence of selective distortion or ridicule. As might be suspected, the evidence presented on this assumption is not overwhelming, although it is reasonably convincing. The possibility that individual differences in drawing content might be important is considered in one sentence. However, to the reviewer, Dennis' use of drawings to explore group values is probably on relatively firm ground without opening Pandora's Box of individual differences, at least for the present.

The technique is as simple as one has come to expect drawing techniques to be: a 8½″ × 11″ sheet of paper, a number two pencil, and instructions to draw a "whole man." The product is then analyzed into categories for which good inter-rater reliability is a reasonable goal: (a) the presence of undesirable characteristics (facial scars, mouth corners down, crippled), (b) traditional *vs.* modern dress, (c) emphasis on masculinity (facial hair, pipes, or cigars), (d) smiling and nonsmiling faces, and others. The relevance of these items for gaining insight into the values of the group is discussed in brief detail. The groups presented for illustration are varied indeed: from the orthodox Hassidim in Brooklyn to middle class boys in Heidelberg, to boys from a village in Cambodia, and to other points in Mexico, Asia, Europe, and the Middle East.

Wayne Dennis has presented a provocative approach to interpreting drawings of human figures. That should be particularly stimulating for cross-cultural and culture change research. While his method may seem to have little value in individual assessment, it does open vistas for changing emphasis in psychological interpretation from the psychopathological to the positive and negative values of the individual. This highly readable volume is heartily recommended to social psychologists, sociologists, and others concerned with intergroup research.

Philip Himelstein
Texas Western College of the University of Texas

The Results of Counseling by D. P. Campbell. Philadelphia: W. B. Saunders Co., 1965. Pp. 205. $6.00

How many individuals secretly wonder whether counseling students really makes any difference in their lives? Here is a book, written in a concise and spirited manner which gives a clear affirmative answer. It deals deftly with a great deal of statistical information, enough to convince the most captious critic. And to frost the cake, it provides reassuring statistics on stability of testing, lack of decline in ability of this generation of college students as compared with those a quarter century ago, improvement of mental ability in adulthood, and sidelights on liberal-conservative temperaments in our society.

The basic material came from files in the counseling office at the University of Minnesota. Data were accumulated from 1933-36 for 384 pairs of students matched on College Board Scores, English skills, age, sex, high school size, and class. In each pair one person was counseled and one was not counseled. In 1962, 731 of these former students were contacted and complete information was collected on 80 per cent of them. Small but consistent differences indicate that over the years counseled students when compared with non-counseled ones are mildly less well satisfied with themselves, but have more accomplishments to their credit. They are less satisfied with their jobs, their marriages, their social life, and their over-all situation. But they have won more degrees, more honors, and more leadership posts. They make over a thousand dollars more per year, and score higher on a gross measure of contribution to society. It is truly surprising that differences such as these can be found after a lapse of a quarter century.

The book contains an excellent chapter on a review of the literature which focuses on motivated control groups, detailing the studies of Berdie, Jesness, Aldrich, Volsky and others, Hoyt, Vosbeck, Williams, Golburgh, Rothney, and Schofield. This chapter admirably brings together material otherwise scattered and hard to find. Some of it involves high school counseling.

The final chapters of the book explore three ancillary issues made available by the data. Each issue is handled with much more detail than can possibly be discussed here. Nevertheless, a few remarks are in order. First, the data indicate that despite a much larger percentage of the population attending college now, there has been no decline in mental ability when 1962 freshmen are compared with freshmen in the 1930's. Second, it is shown that the former freshmen, now adults, score higher on the same tests now than they did then—a finding indicating a continued rise in ability. Finally, there is a fascinating dichotomy of the group into liberals and conservatives. The liberal is likely to be less religious, more unconventional, and have published more. The conservative is more likely to be a religious Protestant and to have a stable marriage.

The volume is well worth reading, for it raises some questions while settling others. Is the "divine discontent" in the counseled student what made him seek counseling in the first place? Is it associated with artistic temperament, and with creativity, which would perhaps help him to write and to publish more? Do conservatives go to counselors and psychologists less because they need them less or because they are less aware of their problems? Is there more mileage to be gained by the counselor for selecting creative, mixed-up students to counsel than the stolid types? These are only a few of the questions which this volume raised for the reviewer. It will pose some equally fascinating ones for others. The reviewer urges the reader to obtain and to study this volume at his earliest convenience.

JOHN C. GOWAN
San Fernando Valley State College

The Outcomes of Counseling and Psychotherapy by T. Volsky, Thomas M. Magoon, Warren T. Norman, and Donald P. Hoyt. Minneapolis: University of Minnesota Press. Pp. 209. $5.50.

A carefully-done ten year study from the University of Minnesota Counseling Center on closely matched students yielded no results when the students were evaluated on anxiety, defensiveness, and problem-solving abilities before and after counseling. The design was to process one student (experimental subject) immediately through brief counseling while delaying his matched pair (control subject), and then giving tests at the start and end of the counseling for the experimental subject (and hence before the counseling of the control subject). The design was clever in avoiding the bias of not counseling a group which wanted counseling experience or of controlling with a group which did not want such experience, but it ran into the negative corollary of Murphy's Law which states: "While riding to the dentist's office for an emergency extraction, your inflamed tooth will begin to feel better."

The authors do a good job in reviewing the field and in setting up the design; they make up for this by a rather poor job when the egg is on their faces at the end. It is very possible that the students were not evaluated on enough instruments, and that the authors were too hopeful of the results of only a few (usually three or more) sessions of counseling to effect measureable change.

This book inevitably suffers by comparison with Campbell's *The Results of Counseling: Twenty Five Years Later* (Saunders, 1965), which curiously reports results from the very same source. In a more deft experiment and write-up Campbell obtained positive results. It is fascinating to think of the political dynamics which sent the authorized version to the university press, and Dr. Campbell's version to Philadelphia. Although it is honest to have negative results

in experiments reported, it looks a little bit as if too many cooks spoiled the broth.

John C. Gowan
San Fernando Valley State College

Vocational Guidance Readings by George D. Demos and Bruce Grant (Editors). Springfield, Illinois: Charles C. Thomas, 1965, Pp. xxiv + 521.

The purpose of this book is to present in an integrated manner the multiple aspects of vocational guidance. The articles were selected as to their appropriateness to the theme of a chapter. No attempt was made to make a given chapter a comprehensive review of a topic. Each selection represents a sample of the typical concerns in a given area. The selections for the book were obtained from the following sources: (a) *Vocational Guidance Quarterly* (24); (b) *Personnel and Guidance Journal* (15); (c) pamphlets, monographs, government publications (11); (d) popular magazines (8); (e) newspaper articles (4); (f) original paper (4); and (g) book (1).

The book is divided into the following chapters: (1) Socio-economic implications for vocational guidance; (2) Vocational development and vocational choice; (3) Industrial, company, and vocational careers; (4) Occupational aspirations; (5) Occupational values; (6) Prestige factors; (7) Work opportunities; (8) Occupational information; (9) Sources of industrial, company, and vocational information; (10) Topics to investigate in studying industries, companies, and vocations; (11) Classifying and filing career information; (12) Evaluating career information; (13) Educational preparation in relation to vocational guidance; (14) Teaching of occupations; and (15) Vocational Guidance. Each chapter has a short introduction which sets the theme for it. In addition each chapter is concluded with a list of suggested additional readings.

The articles represent a mixture of theory and practice. The basic premise underlying this approach is the idea that given a theoretical orientation and an example, the reader could put into practice in a school setting a procedure that is based on a psychological premise. In this book of readings, the execution of theory into practice is at times blocked by the incompleteness of the theoretical presentation, the inappropriateness of the example or the lack of one, or by the sequence of articles. For example, there are twelve chapters between the discussion of vocational development and choice and vocational counseling. There is no mention that these ideas are interrelated and that counseling facilitates development and choice. In addition there is an imbalanced presentation of the theories of vocational choice: Holland's theory receives

a thorough presentation but other theoretical orientations (e.g., Super, Roe, and Tiedeman) receive little if any attention. The chapter on "Classifying and Filing Career Information" has several good examples, but the "Introduction" fails to mention some of the pertinent factors that should be considered when a filing system is established.

There are several areas of vocational guidance which were not discussed in this book of readings. Assessment of behavior is one of them. Some discussion of test procedures, test interpretations, and allied topics should have been made, for they are pertinent to vocational guidance. Although teaching of occupations is presented, other group procedures are not discussed. The organization and administration of vocational guidance programs which is the concern of many counselors and administrators is not discussed. If theory is to be put into practice, some structure is needed to implement it. Consequently, a review of organizational and administrative concerns would seem appropriate.

This book of readings provides only a minimal coverage of the area of vocational guidance. Certain areas like aspirations, values, and occupational information have a very good selection of articles in that they give both depth and breadth. In some of the other aspects of vocational guidance, the range of articles is somewhat limited. In one sense this book could be termed a sampler: it stimulates interest in the field of vocational guidance but fails to satiate it.

HENRY KACZKOWSKI
University of Illinois

Decision Making in the Employment Interview by Edward C. Webster. Montreal, Eagle Publishing Co., Ltd., 1964 Pp. iv + 124. $3.50.

This brief volume is largely a compilation of excerpts from doctoral dissertations conducted at McGill University under the direction of the author who is Professor and Chairman of the Department of Psychology. The volume summarizes nearly a decade of research made possible through grants from the Canadian Defense Research Board to McGill University. Dr. Webster, who is editor of this volume, was the principal investigator of the research supported by these grants.

The settings for the extensive research on the merits and limitations of the employment interview reported in the volume are drawn from both the military and industrial fields of occupation. The principal thesis in these experiments is that the employability of personnel is a complex and important matter of concern to both the employee and employer, and that the successful prediction of success in the employment situation is determined by the skill and

objectivity with which the interviewer is able to perceive, analyze, and integrate the material which is presented to him in the process of the employment interview.

A very complete and revealing review of the literature, ranging from the early 1920's to the present day, on the matter of the accuracy of interview impressions constitutes the major portion of Part I on this report. It is based largely on the dissertation work of Springbett (1954) and Sydiaha (1958). The major conclusion reached by these two authors is that the general consensus of psychologists concerned with the accuracy of the employment interview, from E. L. Thorndike to J. P. Guilford, is that the reliability of the interview techniques leaves much to be desired.

The subjects of the McGill experiments were, in the main Personnel Officers in the Canadian Army and recruits or men awaiting enlistment. The studies reported in this volume, other than those by Webster's students cited above, were conducted by Anderson (1961), Crowell (1961), and Rowe (1960). A bibliography of 133 references on the general topic of the employment interview accompanies the next of this volume.

The volume itself is organized around the dissertation excerpts and unpublished papers on the general issue of decision making in the employment interview by the five principal collaborators of the text. The reports are interesting, although a critical reviewer can find much fault with the experimental design and conclusions of each paper; they tend to be somewhat disjointed and as a whole present a rather fragmented picture of the total problem area being investigated; and the reader is left at the end with a rather confused and somewhat pessimistic impression of what was actually accomplished by the lengthly project. One of the few positive conclusions drawn from these studies is that the interview as a predictor of employability fares much better in presenting negative evidence, i.e. identification of the unsuitable candidate. But then, this situation is the same for all other psychological measurement techniques and devices. Another definite finding was that unfavorable characteristics of the candidate contribute most heavily to the decision of rejection for employment.

The report ends with a list of suggestions offered by the principal investigator. These recommendations do not appear to be entirely forthcoming from the findings reported in the text. Rather they seem to go far beyond what was demonstrated by the McGill experiments, and as Dr. Webster aptly points out in the introduction to Part V "This chapter would have been easier to write had the experiments reported not been undertaken." The reader too is left with this last impression. Had the studies not been conducted and the findings compiled, he might have retained some assurance in the value of the interview as an assessment technique.

REFERENCES

Anderson, C. W. The Relation of Verbal Behaviour to Decision Formulation in the Employment Interview. Unpublished doctoral dissertation, McGill University, 1961.

Crowell, Areta H. Decision Sequences in Perception. Unpublished doctoral dissertation, McGill University, 1961.

Rowe, Patricia M. Individual Differences in Assessment Decisions. Unpublished doctoral dissertation, McGill University, 1960.

Springbet, B. M. Series Effects in the Employment Interview. Unpublished doctoral dissertation, McGill University, 1954.

Sydiaha, D. The Relation between Actuarial and Descriptive Methods in Personnel Appraisal. Unpublished doctoral dissertation, McGill University, 1958

PETER F. MERENDA
University of Rhode Island

Emotions and the Job by S. G. Roog and C. A. D'Alonzo. Springfield, Illinois. Charles C. Thomas Publishers. Pp. ix + 192.

More man-days of work are lost in industry through the crippling effect of emotional disturbance and mental health imbalance on the part of the worker than is generally realized. Drs. Rogg and D'Alonzo document this fact through the scrutiny and appraisal of the records of more than 100,000 employees of a large industrial concern over a period of 35 years. Dr. Rogg is a psychiatrist, and Dr. D'Alonzo is Assistant Medical Director of the E. I. du Pont de Nemours and Company. Both have had wide experience in the diagnosis and treatment of emotional problems of the American worker. In their present volume they share this information with the readers by describing typical examples of the mental health problems of workers, and they attempt to show both the manager and the line worker how to use effectively these data and reports.

To say that the coverage of mental health problems in this volume is extensive is certainly an understatement. A total of twenty-two topics is covered. They range from a discussion of anxiety and fear reactions to the more acute illnesses such as schizophrenia and melancholia. Also included are brief expositions on motivation, intelligence, personality development, and the problems of adolescents. Alcoholism, hypnosis, accident proneness, suicidal tendencies, and the use of drugs by employees constitute a major portion of the remaining topics. Although the coverage is extensive, it is by no means intensive in this brief text. Most of the topics are discussed in single chapters, some of which are only two to five pages long. The reader, therefore, must be careful not to assume that he can acquire much sophistication or knowledge of the many concepts to which he is superficially introduced by the authors. However, to

their credit, the reader is cautioned early by them not to treat the contents of the text as recipes for solving human emotional problems.

Many interesting and timely data are presented in the tables (labelled figures by the authors) of the latter chapters of the text. Particularly relevant and up-to-date are such data as per cent of employees taking psychotropic drugs by age and sex; consumption of alcoholic beverages before dinner, by age; leading causes of death by age groups; and average annual suicide rates.

The treatment of psychological concepts and principles is quite well done for a layman's consumption. It is accurate, precise, and interesting. Perhaps the only lapse in accurate reporting is found in the authors' reference to "Dick" Armstrong rather than Jack Armstrong, the All American Boy, when they are describing the stereotype ideal self-concept.

Managers and employees both are likely to gain some helpful insights into the understanding and solutions of mental health problems through the reading of this informative report.

PETER F. MERENDA
University of Rhode Island

Thinking: From Association to Gestalt by Jean Matter Mandler and George Mandler (Editors). New York: John Wiley & Sons, 1964. Pp. x + 300. $4.95

This is an excellent compendium of thinking on thought which could be used to advantage by any psychologist interested in cognition. The volume, which traces a path from Aristotle to Duncker, includes previously untranslated work. A scholarly and informative running commentary by the editors transforms the volume from a collection of readings to one that might well be used as a textbook in a course on thinking, accompanied by, perhaps, one of the current collections of readings in the cognitive processes, or by *Computers and Thought,* edited by E. A. Feigenbaum and J. Feldman, a collection of research in computer simulation. The latter work builds on this foundation to an extent not ordinarily realized, as the Mandlers indicate.

The table of contents provides an outline to the potential user. The volume starts with early associationism; Aristotle, Hobb, Locke, Hume, Hartley, James Mill, and Bain are sampled. Next, thinking and the imageless thought controversy are covered in the work of Mayer and Orth, Marbe, Messer, Ach, and Titchener. Last, directed thinking and the unit of thought are considered as probed by Watt, Ach, Külpe, Müller, Selz, Koffka, Wertheimer, and Duncker.

The book is handsome and well made. Its price is perhaps a little high ($4.95) when one considers the volume's size. For those will-

ing to give thought a place in modern psychology, this book is highly recommended.

EDWARD J. O'CONNELL, JR.
Syracuse University

A Primer of Experimental Psychology by Joseph Lyons. New York: Harper & Row, 1965. Pp. 322 + xii. $3.50, paperback.

Although psychologists have always been an argumentative lot, the contemporary scene in psychology is one that particularly abounds with controversy. The present generation of psychologists —above all else—are schooled in critical, logical thinking; to them most of the old answers are unsatisfactory and everything is to be questioned. In such an atmosphere it is not surprising to find books being published on controversies within psychology—controversies in learning theory, in perception, and in mental disorders. Another earmark of the current scene is the increasing number of scholarly paperbacks that are being published. *A Primer of Experimental Psychology* has both of the characteristics referred to; it deals with controversial topics in psychology, and it is a paperback.

The author, Joseph Lyons, is a research psychologist at the VA Hospital in Lexington and a lecturer at the University of Kentucky. Among his research interests are social perception and bodily orientation in space, and he is the author of *Psychology and the Measure of Man* as well as the present book. Professor Lyons has a flowing, easy-to-read style, but it is obvious that his "experimental psychology" is not quite the same as the traditional experimental psychology of Woodworth and Schlosberg or Underwood. The experiments which Lyons cites, with only a few exceptions, have more to do with clinical and social psychology than with the "pure" experimental psychology of sensation, perception, learning, and memory. In addition, this book is no *primer*. Before discussing the results of selected investigations, Lyons presents certain background material essential for an understanding of the investigations. But the average beginning, unsophisticated student of psychology will still have difficulty comprehending some of the ideas and arguments. Consequently, the book will be more useful in upper-division and graduate courses than in the elementary courses for which the publishers reportedly intend it.

The book, one of a series on experimental psychology planned by Harpers, consists of eight chapters: (1). The Nature of Experimental Inquiry; (2). An Introduction to What Experimenters Do; (3) Specifying the Experimental Problem; (4) The Logic of Instruments; (5) Experimental Control; (6) The Formal Characteristics of Experiments; (7) The Problem of Awareness; and (8) Frontiers in Experimental Psychology. Throughout the eight chapters there is an emphasis on methodology, contemporary issues,

and the philosophy of scientific psychology. The book is not a collection of findings or results of experiments. Rather, the author has selected examples of investigations concerned with certain controversial topics and has organized the text around these topics. The six headings of Lyons' "Library of Controversy" sub-sections are: Research and Psychotherapy, Sight and Space, Race and Intelligence, Mazes and Learning Theory, Experiments on ESP, and Learning and Awareness.

In summary, this is a brief, interesting, well-written book on the factors which go into designing, conducting, and interpreting the results of experiments in psychology. Lyons gives an overall view rather than spending much time on particular matters, and the examples which he employs as illustrations of specific ideas and methods are uniformly captivating ones. The book will be useful as supplementary reading for students who have some background in psychology and for the professional psychologists, but it is not recommended for the tyro.

Lewis R. Aiken, Jr.
Guilford College

Learning and the Educational Process by J. D. Krumboltz (Editor). Chicago: Rand McNally and Company, 1965. Pp. xiii + 277.

In the summer of 1964 a conference was held at Stanford University with the aim of stimulating research on educational problems. Support for the conference came from the Cooperative Research Branch of the U.S. Office of Education and the Carnegie Corporation. The idea for the summer program was generated by a Social Science Research Council Committee which had as its goal stimulation of basic research on educational learning. From the many who applied only 40 young researchers could be chosen to attend. Fortunately for the rest of us who were not there, a written record of ten conference presentations has been made available in the present book.

Learning and the Educational Process was compiled with the same goal that motivated the conference; that is, to stimulate research on educational problems. It is intended that the book serve as a resource for instructors and their graduate students, for professional educators, and for researchers in fields related to education.

A review of the general topics covered in the volume reveals three areas of emphasis: research methodology, motivation, and language. Three chapters have as their theme the methodology for research on educational problems. Robert Gagné argues for the statement of educational objectives in terms of human performance. Lawrence M. Stolurow suggests the use of a teaching model for

investigations into problems of school learning. His comments about the distinctions between descriptive and predictive models seem particularly appropriate. In a chapter titled "School Learning over the Long Haul" John B. Carroll asks for longitudinal research into the long-term effects of curricula. He identifies aptitudes, perserverance (motivation), opportunity to learn, the quality of instruction, and ability to comprehend instruction (language) as relevant variables needing longitudinal study.

John W. Atkinson, Daniel E. Berlyne, and Jerome Kagan write about their particular areas of concern within the general topic of motivation. Atkinson presents a clear statement of the effects of need for achievement and fear of failure on human performance. His work is an example of the methodological approach about which Gagné writes. The role of conflict in arousing curiosity drive and of various methods of presenting information to reduce this drive is discussed by Berlyne. It is worth noting here that at several points the authors of the various articles seem to be in disagreement. Atkinson argues persuasively against the use of drive reduction theories in education; Berlyne uses one. Kagan's contribution is concerned with the differences in behavior of children identified as impulsive or reflective.

Ernst Rothkopf's article on ". . . Problems in Written Instruction" has more general interest than its title might indicate. It bridges the gap between the "motivationalists" and those whose primary concern is with language. It is, along with Atkinson's chapter, a good methodological model for the educational researcher. The major focus of his concern is the behavior involved in successfully learning from written material, something which Rothkopf calls "mathemagenic" behavior.

The positive effects of middle class homes and the negative effects of lower class homes upon the language development of children is discussed by Strodtbeck. Loban echoes this theme in a brief report of the findings from his longitudinal study of the language proficiency of school children. Evan R. Keislar and Larry Mace report on their research into the effects of the sequence of speaking and listening training upon learning in foreign language.

This is a brief overview of the ten conference presentations. The general competency of the work presented is high. The reviewer believes that the book will serve well its purpose of generating better research into educational problems. Though the sampling was not chosen to be representative, it is instructive to observe the areas of research which these authors are pursuing. Concern for sound methodology and research into language and motivation represent the major themes of this book.

The editor of this volume must be commended for choosing the articles presented here. However, it must also be said that no at-

tempt is made to provide a structure for the overlap and diversity in the book. Occasionally, one author referes to the work of another within the book, but this seems less integration than afterthought. Krumboltz hopes that the dissonance produced by the diversity of views "will result in constructive efforts at resolution." The constructive direction of dissonance resolution could have been made more likely, however, by the introduction of a summary chapter, or, perhaps, the exchange of comments by the authors on one another's papers.

Learning and the Educational Process or, what would have been a more appropriate title, "Research and the Educational Process," hits the mark. It is provocative reading for those who identify, or would like to identify, educational research as their working domain.

LOREN S. BARRITT
The University of Michigan

School Psychology by Roger Reger. Springfield, Illinois: Charles C. Thomas, 1965. Pp. xv + 213. $7.75.

Within the last two decades the school psychologist has emerged as a significant figure in the educational system. Unfortunately, the profession, in its evolution, has never had a well defined role within the educational system. Because of this, educators frequently question the value of the school psychologist's services and relegate the psychologist to the role of test administrator and reporter.

Roger's book represents a recent trend on the part of several concerned individuals to clarify the aims and goals of the professional school psychologist. The book is the work of a practicing school psychologist who has been taking some very critical looks at the practices in the profession. This worthwhile, though difficult task is accomplished by discussing a number of contemporary problems, practices, and issues in the field which have hindered the development of the profession.

The book is divided into three main parts: (a) the roles and identity of the school psychologist, (b) issues in school psychology, and (c) a theory of special education. In part one Reger sets forth his major argument that the school psychologist should be considered an educator, by first pointing to the present inadequate role of the psychologist and then by elaborating on the new roles and identity the profession should seek. He contends that the contemporary emphasis of the profession on pathology, its identification with clinical psychology, and particularly its psychometric orientation have lessened the effectiveness of the psychologist and his contribution to the education of the child.

"Educational philosophies and practices continue to change, but

school psychologists have put themselves in the position of being followers, rather than leaders, of the processes of change. Psychologists, trained in theory, practice, and research, probably will continue for some time to come to administer their tests, score these tests, and smugly interpret the results to receptive school personnel. And while psychologists continue to talk about score patterns, ego boundaries, and superego conflicts, 'real IQs', 'potentials', 'prevention', and a salad of other meaningless nonsense, change will take place all around."

Reger believes that a more meaningful identity for the school psychologist would be in a model which resembles the college academician, enabling the psychologist to function as an educator rather than a clinician. The primary function of the psychologist within the role of educator would be one of planning educational programs for children. With the psychologist placed high in the administrative structure, he would have the flexibility to function, not in the narrowly prescribed role of psychotherapist or test administrator, but as a teacher, programer, and researcher. The psychologist's basic purpose in the public school system would be to improve the quality and effectiveness of the educational process through the use of psychological knowledge. In essence Reger contends that the school psychologist should be viewed in the role of an educational practitioner as well as a scientist utilizing the latest methodology and techniques in the design of educational plans.

Although not all readers of this journal would agree with Reger's view of psychological tests as being "primitive instruments" which secretaries can administer and interpret, one is impressed with his attempt to conceive of the psychologist, not as a technician, but rather as a professional expert with one of many specialties being in the area of psychometrics.

Part two, which constitutes almost half of the book, is devoted to contemporary issues within the profession, including a discussion of (a) the predictive validity of IQ tests, (b) the lack of much needed research by school psychologists in their profession, (c) the value of psychotherapy in general and the question of the necessity for therapy in the schools, (d) the failure of the "medical model" in diagnosis and the need in diagnosis for information more relevant to educational planning, (e) the "openness of information" to parents in psychological reports, and (f) the interaction between the psychologist and teacher.

The issues raised by Reger are not altogether new and some, e.g. working conditions, tend to distract from his main arguments. Although the question of the validity of IQ tests has not been adequately resolved, it is debatable whether the validity of IQ tests "has hardly been noticed" by psychologists. In a similar vein, downgrading psychological tests in general without elaboration or

clarification may tend to confuse and prejudice the lay reader, particularly since Reger also points to the necessity for psychologists to be well trained in psychological testing.

Part three of the book outlines an educational theory for exceptional children and stresses the need for a theory which considers the needs of the individual child rather than builds programs around "classifications" of children presumed to have the same characteristics. The ideas presented here are worthy of consideration, but unfortunately the inclusion of this section does not add to the author's main thesis, namely defining the goals and functions of the school psychologist. The impression one receives is that the author has much to say about school psychology, educational philosophy, and professional competency, and for the sake of expediency has placed them all in one book.

This book reflects more the author's own experiences and views than the documented opinions of others. As such, the references are not integrated very well into the chapter discussions, but they are quite substantial and extensive. The chapters within each section are organized well, but at times they tend to be redundant. For the most part, the language which is nontechnical should not pose difficulties for lay people.

Throughout the book, Reger has taken a very critical view of the entire profession of school psychology, and a number of psychologists reading it will no doubt disagree and even be irked at his analysis of many of the present practices in the field and by his suggested role for the school psychologist. It is a book, however, that should have been written years ago. The author has made a sincere and worthwhile effort to clarify the structure and the goals of the professional school psychologist. As such, Reger has made a significant contribution. Although this volume cannot be considered a textbook in the usual sense, educators, practicing psychologists, and graduate students would find it worthwhile to read and consider carefully.

RALPH B. VACCHIANO
Fairleigh Dickinson University

Psychology and Education of the Gifted: Selected Readings by Walter B. Barbe (Editor). New York: Appleton-Century-Crofts, 1965. Pp. x + 534.

This book of readings on the psychology and education of the gifted consists of 54 papers, primarily journal articles, dealing with philosophical presentations, discussions of different points of view, reviews of literature on specific topics, and significant research reports. Most of the papers were published before the sixties; the editor points out that "inclusion was determined by importance in the field rather than recency of publication."

The editor has attempted to bring together a collection of papers which present significant, and sometimes contradictory, points of view; he has done remarkably well in collecting much of the outstanding literature on the psychology and education of the gifted. While it is recognized that no book can be expected to include all the relevant work in the field, Barbe has collected readings which sample the field quite adequately.

The material has been organized in such a way that the readings are meaningful. An introduction by the editor precedes each of the four major parts and provides the continuity needed for such a book. Part I presents an introduction to the study of the gifted by devoting sections to historical development, concepts and concerns about giftedness, and facets or traits of mental giftedness. Appropriately, the pioneer work of Terman, Guilford, and Thurstone is included in this section. An obvious omission is the work of Hollingworth.

Part II, Measurement of Giftedness, includes three articles on identification and testing the gifted; three articles were selected on non-intellectual factors and giftedness.

Articles which describe the characteristics of the gifted are included in Part III, Background of the Gifted. The first four papers in this section form a logical section in which hereditary factors and family background are considered. A section is also included on the social and emotional characteristics of the gifted.

Part IV includes five sections which concentrate on the development and encouragement of giftedness. The emphasis in these sections is upon issues in the education of the gifted, factors in motivating and developing the gifted, problems related to good adjustment, and ways to plan for the gifted in the elementary school and in the high school.

Part V includes more issues in the area and a section on research. Although evaluation is not mentioned in this section, the last article presents detailed procedures for carrying out an evaluation. The article also gives a good example of an evaluation by describing how a school system went about the task of evaluating the effects of segregating the intellectually gifted pupils in homogeneous special-progress groups on the junior high school level.

As the editor has stated in the Preface, this book of readings is designed for the advanced student. While each of the articles taken separately provides for stimulating reading, the editor has attempted to arrange and combine the book for maximum clarity and meaningfulness. An outstanding characteristic of this volume is that it provides important samples of the thinking which has been influential throughout this century.

The reviewer has one reservation about this book regarding the relatively small number of articles on the measurement and identi-

fication of giftedness. Inasmuch as the assessment of giftedness represents a major area of concern to both psychologists and educators alike, a greater emphasis should be devoted particularly to the measurement of higher mental processes of the gifted student.

Bert W. Westbrook
North Carolina State University

Teaching the Culturally Disadvantaged Pupil by J. M. Beck and R. W. Saxe (Editors). Springfield, Ohio: Charles C. Thomas, 1965. Pp. xvi + 335.

Less than a decade ago the rising tide of delinquency urged upon the government the need to promote research of and training for unemployed school dropouts. Dropouts, it was believed, provided the soil upon which delinquency breeds. In addition to stop-gap measures, the Kennedy era ushered in developmental programs for potential dropouts. The first programs were begun in economically deprived communities and neighborhoods, where ignorance, unemployment, and inertia were believed to flourish. The jeremiad Michael Harrington bruised our complacency and made us stand up and take notice of the mores, hopes, and disenfranchisement of a neglected fifth of the population.

Available now are the records of university based endeavors as well as of public-education based undertakings such as those in Philadelphia, New York, St. Louis, and San Francisco. The data do appear to be helter-skelter. In all fairness to the pioneer workers, with little theory as a guide, a purely empirical approach was justified and salutory. The time has come, though, to pause and review, synthesize, and select the most feasible approaches to educating those from humbler circumstances. Schools of education are preparing to increase their efforts in developing a pedagogy for the poor. The outstanding school administrators have insisted that their teachers take in-service courses on the disadvantaged. The book by Beck and Saxe meets a demanding market.

Will the book satisfy this market? Today's knowledge about "the disadvantaged" may be likened to human prenatal development. No longer a zygote, the organism is maturing into a fetus. The educational mother is just beginning to suspect that she is pregnant. The Beck and Saxe exorcising of the fetus at this time reveals in four chapters what are the basic characteristics of the disadvantaged, in eight chapters what the elementary school could do for him, and in three chapters what the community could do for him. The impression this reviewer has is this: The differences between the fetus under discussion and that of any other species is hardly recognizable. Several of the 17 contributors to the book acknowledge and are even apologetic about the modicum of viable knowledge they impart. Even as they awaited their galley proofs,

cell division was constantly taking place. Information was taking on new structures. While the stethoscope the editors applied to the disadvantaged might have detected the heartbeat of the fetus, it may only have been the mother's heartburn.

The spirit which guided this book is to be praised. The tone set in Havighurst's introduction is that of dedication and commitment. The hope is to raise the intelligence quotients of low income families by ten points, to eliminate mental retardation due to "environmental deprivation," and to remove more than half of the academic retardation in elementary school. This is Havighurst's five-year plan. The information presented in the book concerning low income families is prolific, although of uncertain validity and reliability. This condition beseeches thinkers to organize the contributors' observations. This call is answered by Brottman in chapter one. He posited a "social system" input and behavior output model. The social system bifurcates into nomothetic and idiographic lines. This reviewer is not certain that the chain of inputs like institution →role →role expectation are causally related. Nevertheless, a model is available. In chapter four, Hirsch postulates a different model but with many elements in common with that of Brottman's. Still later, a communication model is presented. Although the multiplicity of models intrigues the research worker, it may confuse the teachers who are to be the main readers of this book. The book which leaves the teacher uncertain about which model to follow in instruction fails in its mission to educate.

The title of the book prepared this reviewer for a description of a population quite different from other youth. Not only this, but it orients one to expect a discussion of the unique techniques tailored to the proclivities of this unique population. This reviewer searched for the distinguishing characteristics of this population of disadvantaged. While clues are given that they may be found among Negroes, Mexicans, inhabitants of rural America and working class homes, what attributes common to these groups might one look for? One writer stated, though, that the majority of people from working class homes give their children a good start in life. Several authors maintain that the disadvantaged do have mores and folkways different from others. Among the disadvantaged, the attitude toward school is negative. Their children play infrequently with toys or with children. Reading matter is absent from their homes. Among the behavioral attributes identified by the writers as distinguishing were inferior auditory discrimination, poor learning habits, and inadequate reading. One may ask what kind of auditory discrimination labels one "culturally disadvantaged," but another as inadequate in auditory discrimination? It seems that too little attention has been paid to the validation of the construct of "the disadvantaged," to the development of yardsticks to meas-

ure it, and to the formulation of an extensional definition of it that would be useful to any external observer. If the functional relationship between change in such attributes as reading ability with remedial treatment is no different for a group "labeled" disadvantaged than it is for those without the label but with the same ability and with the same treatment, there is little reason to retain the label.

The chapters dealing with teaching of arithmetic, language, social studies, art, and music are particularly valuable to a teacher. The games, exercises, and projects emphasize body involvement. The projects range from writing Haiku to pen pal clubs for potential dropouts. Many of these exercises are recommended for those students who do not bare the badge of "the disadvantaged." From the tone of these chapters this reviewer felt that the authors were describing youth who were *at a disadvantage in school* in reading, in arithmetic, or in science. Whether the disadvantage is due to cultural forces, due to the range of stimuli which a family permits its children to experience, or due to a set of school standards alien to some youth is debatable.

Despite its limitations, the book does contain valuable messages to the educator. From the book one may see how vital is concept learning, how important it is to teach the child to learn how to learn, and what is the motivational value of giving a child something that is his own. Some other salient points were to acquaint present oriented youngsters with the past and to learn the world from the point of view of the child.

Redundancy abounds in the book. The works of Bernstein on public and private language, of Deutsch on concept development, and of Clark on self-fulfilling prophesy are cited in different chapters to fix the same points. In the chapter on language arts this reviewer hoped to see the pedogogical implications of Bernstein's works. This was one of the chapters, however, in which his work was not cited. In conclusion, thinking about the disadvantaged is still inchoate. Many of the novel instructional aids described merit classroom trials. In addition to teaching youth at a disadvantage in school, it would be well to *learn* from them during this next five year period. Hopefully, the editors will examine the embryo.

NORMAN M. CHANSKY
North Carolina State University

Reading Comprehension for Scientists by Richard H. Bloomer. Springfield, Illinois: Charles C. Thomas, 1963. Pp. xiii +213. $8.75.

The need for materials and practice exercises which will help persons to improve comprehension of scientific materials is apparent. If it could be demonstrated that such materials were effective, they

would probably be widely used. It is on this point that *Reading Comprehension for Scientists* must be questioned.

This book of practice exercises is based on passages taken from science texts from a broad range of subject matter and arranged in "levels" of difficulty. Every tenth word is deleted in each paragraph, and these deleted words are dropped to the end of the paragraph in scrambled order. The Key in the back of the book gives the correct order of these words and a "critical score." If the user does not reach this critical score, he will do the next passage of parallel difficulty; if he reaches this score he will move up to the first exercise in the next level.

The author indicates in the index that this approach to practice will teach a student to attend to what he is reading, and teach him to postpone his conclusions until he has read the total context, and thus he will learn the "elemental scientific concepts and the scientific language pattern" (p. vii).

The materials have face validity. Coming from a wide range of textbooks, they are concerned with a large number of scientific topics. However, further questions of validity must be raised.

All exercises are based on the cloze technique. The author does not state how the choice for deletions was made. There are two usual ways; the so-called structural and the lexical. There has been some research to indicate that cloze tests based on these two techniques do not measure the same thing. But a more basic question here is whether the cloze technique used in exercises to teach comprehension skills are valid at all. It seems that the question cannot be answered directly. The author presents no evidence by which the use of this book of exercises can be evaluated and the literature contains a paucity of such evidence. One must, therefore, turn to evidence which has been presented to validate cloze tests.

The most impressive report this writer has found is that done by Weaver and Kingston (1963). On the basis of a factor analysis of reading comprehension tests, modern language aptitude tests, and a scholastic aptitude test, the cloze tests did not show a significant relationship to tests of reading comprehension, but rather to what the authors have named rote memory and flexible retrieval.

Certainly after trying the exercises in this book and watching other adults try them and either becoming discouraged and quitting or approaching them as one would with a puzzle, the reviewer questions whether enough motivation to complete the book could be engendered. (It should be explained that possible right words are supplied at the end of each passage in scrambled order. One trouble with this procedure is that they do not always appear on the same page with the passage. Also by the time one finds any reasonably sensible word to place in a blank, he must go back and reread because he has lost the train of thought developed earlier.)

Therefore, it would seem that the following questions need to be answered before this book is used as the basis of a program purporting to develop reading skills necessary for comprehending scientific material.

1. What skills are necessary for successful reading of scientific material? (Theoretically they differ only in emphasis from those necessary for reading other types of material.)

2. Does the cloze technique applied to scientific material develop these skills (after they are defined)?

3. Will persons in need of developing skills to improve the reading of scientific material apply themselves to this set of exercises long enough to learn these skills?

One other concern. The retail price of this book is $8.75. Students need not put their answers in the text, but an answer booklet should be made available if the book is to be reused. A soft cover would reduce the cost of the text somewhat. Surely potential users will consider cost when sampling texts to recommend to students.

FRANCES TRIGGS, *Chairman*
The Committee on Diagnostic Reading Tests, Inc.
Mountain Home, N. C.

REFERENCE

Weaver, Wendell W. and Kingston, Albert J. A Factor Analysis of the Cloze Procedure and Other Measures of Reading and Language Ability. *The Journal of Communication,* 1963, 13, 252–261.

A Guide to Evaluating Self-Instructional Programs by Paul I. Jacobs, Milton H. Maier, and Lawrence M. Stolurow. New York: Holt, Rinehart and Winston, Inc., 1966. Pp. ix + 84.

The title of this book is somewhat a misnomer. It might more appropriately be called *Understanding, Selecting, and Evaluating Self-Instructional Programs.* The first third of the book is an explanation of what programmed instruction actually is, the second third is devoted to how to select a program, and the last third to evaluation. It must be emphasized that as stated in the preface, this book "may profitably be read by teachers, principals, curriculum specialists and other school administrators who may be unacquainted with this area." The writers might have also added (but did not) "and not so profitably read by psychologists, educators, or any other scholars." This book is not recommended for the typical reader of EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT; he would find the first two-thirds of the book exceedingly fundamental and the latter third mildly interesting. As an example, the table of contents lists two-thirds as many titles as there are pages in the book with such titles as "What Do Programs Look Like," "On What

Educational Principles is Programmed Instruction Based," or "What Types of Response Do Programs Require the Student To Make." In a sense, *this* book is programmed, but there are not directions in the front for certain people to "skip to page 43"! This writer is convinced, however, that the text is certainly a guide for those not in the educational and psychological measurement area. It could prove very handy for a teacher of Spanish who wished to understand, select, and somewhat superficially evaluate a program for her area.

The volume is almost totally functional with little time spent on the theoretical or empirical research aspects. For example, the difference between branch and linear programming is explained almost entirely mechanically with no tie-in to theoretical concepts. Names of famous people working in the area, references to research conducted, films available, names of sources of instructional programs are sprinkled throughout the book. In view of this, it is obvious that the use of the word "Guide" in the title is an appropriate word. It is somewhat lamentable, however, that a greater attempt to insert some thought-provoking material was not made. The latter third of the book does provide some basic information on sampling problems, the need for some basic statistics, and the like. Toward the end of the third part, an actual evaluation study done with the Denver Public Schools is presented. If, however, a piece of research was contemplated by a reader involving the usual statistics, further help would be needed.

In conclusion, this short contribution is definitely recommended as an excellent source of fundamental information on understanding, selecting, and evaluating programmed instruction materials, particularly for the uninformed.

PHILIP S. VERY
University of Rhode Island

The Hidden Dimension by Edward T. Hall. New York: Doubleday and Company, Inc., 1966. Pp. xii + 193.

Hall's very interesting and readable work is a functional combination of the Sapir-Warf hypothesis and social pathology. Beginning with the research work by C. R. Carpenter, John B. Calhoun, and others, Hall extends into the realm of human metropolitan life this basic finding: the crowding of rats in a limited amount of space yields abnormal patterns of behavior that can lead to possible extinction of the entire population.

Obviously, it takes a considerable degree of generalization to go from these experimental laboratory conditions to discussions of inter- and intra-cultural relationships. Hall does this extrapolation on the basis, primarily, of personal experience, with some sup-

port from sociological studies. He is presenting, in effect, an hypothesis:

"The implosion of the world population into cities everywhere is creating a series of destructive behaviorial sinks more lethal than the hydrogen bomb."

This sink is caused by variables ignored by government and society; i.e. the number of square feet per person and limitation imposed on his perceptual space. When limited, man is uncomfortable and antagonistic. The variables which help to control man's emotions include his visual, auditory, kinesthetic, olfactory, and thermal perceptions of his geographical life space. These factors, which are culturally learned, differ from culture to culture—and perhaps more importantly in the United States, from sub-culture to subculture.

If, as Hall states, the population density in already overcrowded cities is to be ameliorated, architects must consider these cultural values in building life spaces more conducive to ecological harmonious relationships. This is not now being accomplished, and, the social upheaval being felt in Los Angeles, Chicago, Cleveland, and other cities may be, in part, due to the ecological disparity between man's cultural heritage and his physical environment.

"It is a mistake of the greatest magnitude to act as though man were one thing and his house or his cities, his technology, or his language were something else." Hall has written this book, it seems, to call for operational research activity to study the elements involved in man's life space. Quantitative measures of such variables are not in evidence, and the current attempts at solutions through Headstart or Wingspread Programs tend to deal with the result, not the cause. As potentially valuable or helpful as such programs might be, it is essential that long-range programs and machinations be developed which can study, measure, and develop large scale city plans accounting for "The Hidden Dimension."

This reviewer found this book upsetting. It explores areas of human behavior which tend to be ignored. If, however, any significant dent on the "culturally deprived" is to be made, or if peoples from different environments and cultures are to be welded successfully (nonantagonistically), this book cannot be ignored. It should be required reading for all educators operating in the field of the culturally disadvantaged, or the culturally different.

Richard E. Spencer
University of Illinois
Champaign